experts | evolving | energy

everoze

AGGREGATED DATASETS – METHODOLOGY

A REPORT FOR WESTERN POWER DISTRIBUTION

CLIENT	Western Power Distribution
DOCUMENT NO.	WESTERN002-AD-R-02
REVISION	B
ISSUE DATE	10 September 2020
STATUS	Final
PREPARED BY	Freya Espir
CHECKED BY	Nithin Rajavelu, Daniel Bacon
APPROVED BY	Felicity Jones

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DISCLAIMER

VERSION HISTORY

A	Initial Draft	28 August 2020
B	Final	10 September 2020

EXECUTIVE SUMMARY

FutureFlex is a joint project delivered by Western Power Distribution (WPD), Everoze and Smart Grid Consultancy (SGC), funded by the Network Innovation Allowance (NIA). FutureFlex is trialling a new DSO service suitable for domestic flexibility, called Sustain-H. This service places data optionality at its heart.

The role of the *Aggregated Datasets sub-workstream* is to quantify the impact of lower portfolio data quality for WPD's DSO services, which will ultimately help to inform payment mechanisms for Sustain-H.

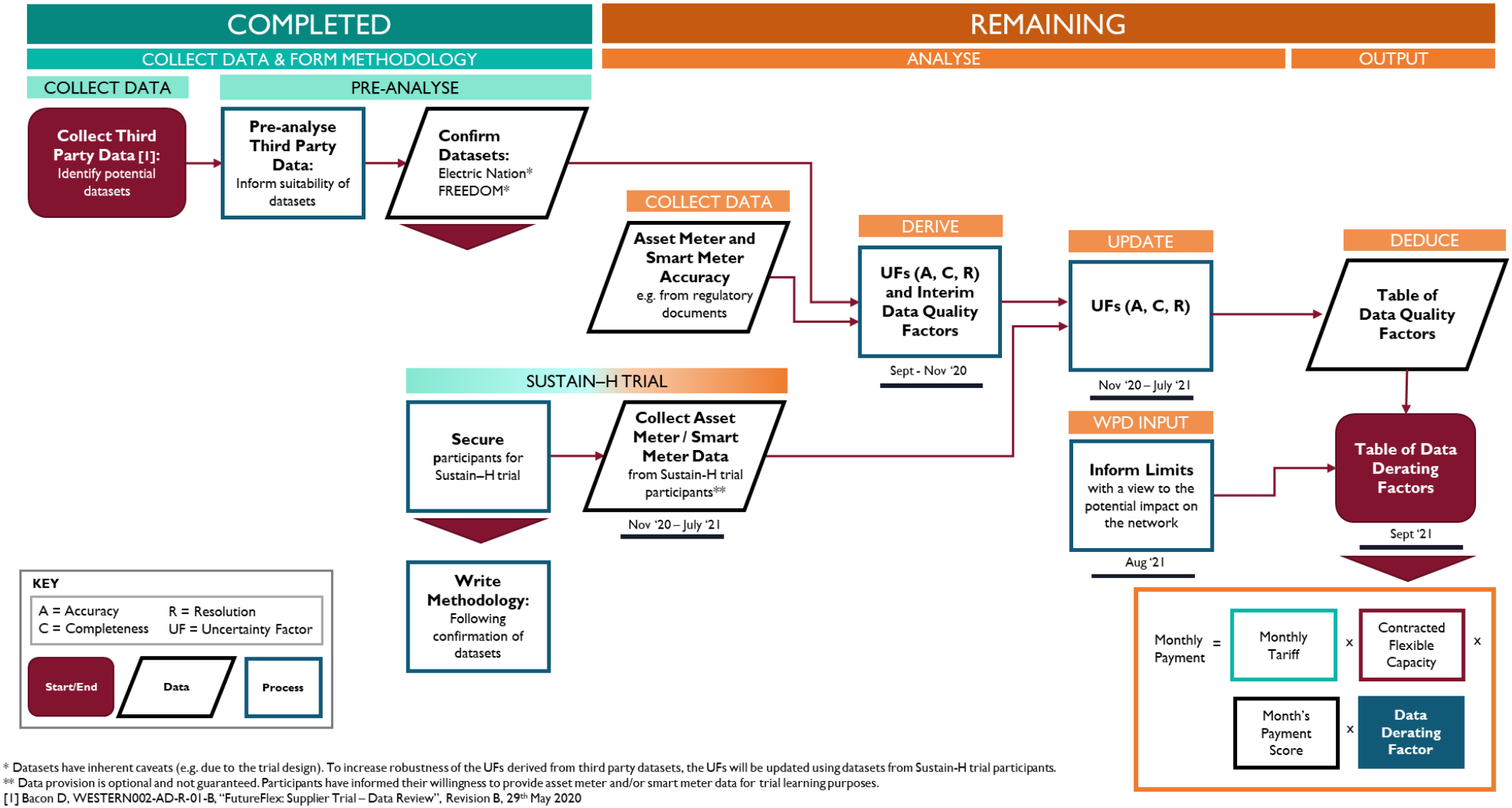
Typically lower dataset resolution, completeness and metering accuracy leads to greater uncertainty in demand prediction. This Methodology Report provides an approach to quantify the impact of these three Uncertainty Factors (UFs) on the measured demand. The UFs will first be defined independently and later combined into a single Data Quality Factor (DQF), which will inform the Data Derating Factor (DDF) for Sustain-H remuneration payments.

To provide a robust approximation of the DQFs, the analysis will first be conducted using third party data from previous trials and later updated following retrieval of smart meter and asset meter datasets from the Sustain-H trial.

Key Outcomes:

- Following substantive pre-processing of two third party datasets: Electric Nation (CrowdCharge) and FREEDOM, Everoze is confident that the data is sufficient for the purposes of estimating DQFs, at least for preliminary purposes
- However, this requires some assumptions around the data including:
    - The power is constant between timestamps for the resolution provided
    - The datasets are representative of real-life usage
    - The conditions within the datasets capture the full variation of usage profiles required to fully assess the DQFs and the DDFs respectively.
- A relatively simple model for defining DQF has been established, through analysis of the resolution, completeness and accuracy of applicable datasets.
- A methodology for determining the inputs to the DQF, which is technology agonistic, has been established and can be applied to pre-trial (Electric Nation and FREEDOM) and Sustain-H datasets.

A flow chart of the *Aggregated Datasets sub-workstream* process is displayed in Figure 1, which captures how the DQF is expected to integrate with the Sustain-H payment mechanism.



* Datasets have inherent caveats (e.g. due to the trial design). To increase robustness of the UFs derived from third party datasets, the UFs will be updated using datasets from Sustain-H trial participants.
** Data provision is optional and not guaranteed. Participants have informed their willingness to provide asset meter and/or smart meter data for trial learning purposes.
[1] Bacon D, WESTERN002-AD-R-01-B, "FutureFlex: Supplier Trial – Data Review", Revision B, 29th May 2020

FIGURE 1: AGGREGATED DATASETS SUB-WORKSTREAM: PROCESS

2. INTRODUCTION

2.1 CONTEXT

FutureFlex is a participant-led trial of second-generation DSO services, deploying step change innovations for procurement, testing and delivery suitable for domestic scale assets. It is a joint project delivered by Western Power Distribution (WPD), Everoze and Smart Grid Consultancy (SGC), funded by the Network Innovation Allowance (NIA).

Workshops held during Phase 1of highlighted a number of challenges for domestic flexibility in meeting DSO data requirements. WPD imposes stringent data quality requirements for its DSO services procured under the Flexible Power platform. Domestic flexibility solutions cannot always easily meet these requirements; for instance, because this adds cost which is prohibitive for small assets. More specifically, workshop participants highlighted the following:

1. Data quality issues were raised as a priority concern for domestic flexibility, culminating in 18 comments on data barriers;
2. Domestic properties tend to produce lower quality of data for flexibility services than larger assets; and
3. The current minute-by-minute metering requirements for DSO services are onerous for domestic flexibility.

In response to this feedback, the FutureFlex team is trialling a new DSO service suitable for domestic flexibility, Sustain-H. This services places data optionality at its heart. The role of the *Aggregated Datasets* workstream is to quantify the impact of lower data quality for WPD.

2.2 OBJECTIVES

The objective of the *Aggregated Datasets* workstream is to help quantify the impact of lower data quality, and thereby inform Sustain-H payments. Specifically, the workstream aims:

- To quantify to what extent aggregation might address data quality challenges at domestic level;
- To establish the implications for DSO service procurement; and
- To provide necessary information to, and integrate with, the Sustain-H service.

The output will be a method to estimate the Data Derating Factor (DDF) within the Sustain-H payment formula.

2.3 REPORT STRUCTURE

The report adopts the following structure:

- **Section 3 – Interface with Sustain-H:** This section introduces the context of the *Aggregated Datasets* sub-workstream which includes the workstream deliverables, summary results from the first deliverable [1], and an introduction to the DDF.
- **Section 4 – Approach:** This outlines the priority methodology objectives for the Sustain-H workstream.
- **Section 5 – Methodology:** This outlines the proposed methodology for analysing data, both from pre-existing datasets, as presented in [1] and Sustain-H trial datasets.
- **Section 6 – Next Steps:** This outlines the next steps for the aggregated datasets sub-workstream, including further research recommendations and a timeline for the analysis.

The first deliverable in *Aggregated Datasets*, "FutureFlex Supplier Trial – Data Review", examined available datasets to inform Everoze's analysis [1]. This report follows on, outlining a methodology.

3. INTERFACE WITH SUSTAIN-H

3.1 APPROACH

***Aggregated Datasets* will develop an approach to calculating the Data Derating Factor (DDF) within the Sustain-H payment formula.** The formula is presented in Figure 2 below, marking where *Aggregated Datasets* fits in.

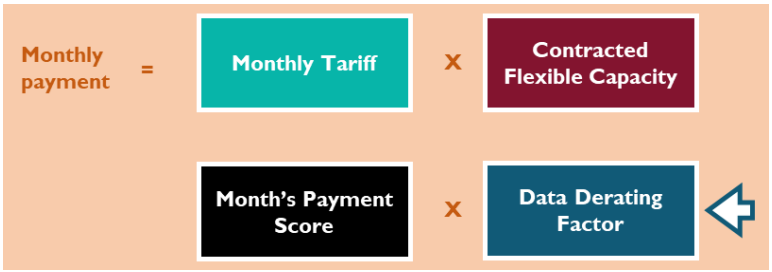


FIGURE 2: PAYMENT FORMULA FOR SUSTAIN-H TRIAL

***Aggregated Datasets* supports Sustain-H by analysing the impact of data quality (attributable to data resolution, completeness and accuracy) for a portfolio of 'N' consumers on the value of the flexibility service.** This is summarised in Figure 3 below. Specifically, it should be noted that:

- Poor quality data is expected to result in an underprediction of demand rather than an overprediction. WPD must be able to predict the actual demand from available data with high enough certainty, and are particularly concerned with ensuring demand is not underpredicted, given the context of network constraint. Typically the poorer the data resolution, completeness and accuracy, the greater the uncertainty in the demand prediction.
- The DDF is a function of 3 Uncertainty Factors (UFs): Resolution Uncertainty Factor (RUF), Completeness Uncertainty Factor (CUF) and Accuracy Uncertainty Factor (AUF). Upload frequency is not considered because Sustain-H is a scheduled pre-fault service, there is no requirement from WPD to provide a real-time dispatch signal to the consumer (or asset).
- The DDF is dependent on selected metering and data gathering solution and portfolio size. The factor is assumed to be independent of technology (EV, BESS and heat pump).
- The DDF is a commercial reflection of the uncertainties quantified. The data analysis as part of this work will quantify the technical uncertainties for each of the three parameters. Ultimately the "cost" of this uncertainty will need WPD input to translate into a DDF.

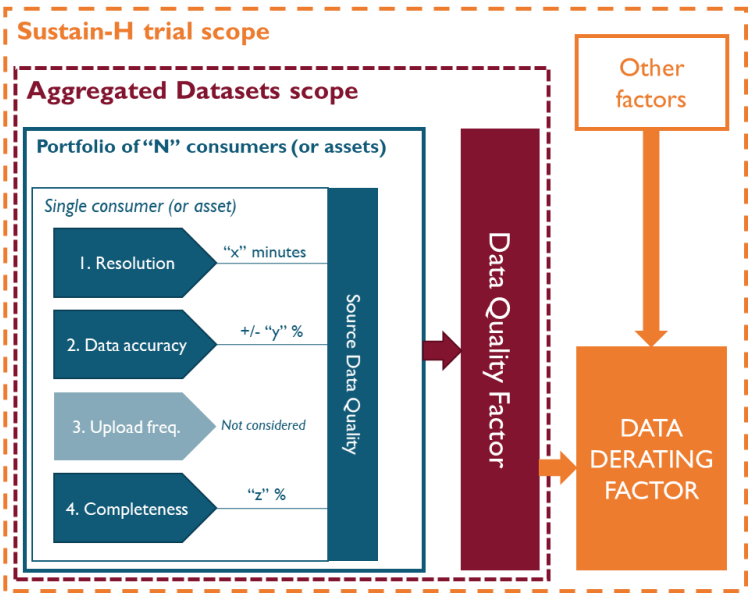


FIGURE 3: PROCESS TO DEDUCE THE DDF

3.2 USING SUSTAIN-H DATA

Our original plan was to use third party datasets to conduct *Aggregated Datasets* analysis. Third party datasets were screened for suitability, as described in Deliverable [1]. The Electric Nation and FREEDOM datasets were downselected as the only suitable options:

- The Electric Nation project [2] by WPD ran three EV smart trials involving 673 households, to investigate the impact of EV charging at home on electricity distribution networks. CrowdCharge supplied one of the two smart charging systems in the trial; these chargers recorded data for each 'charging event' (i.e. transaction) including charging start time, stop time and total charge (kW). Where this report states "Electric Nation", it is only the CrowdCharge dataset under consideration.
- The FREEDOM project [3] by PassivSystems trialled the operation of hybrid gas boiler – Air Source Heat Pump (ASHP) heating systems in 75 homes. Data supplied from the FREEDOM project was of the resolution: 30 minute, 5 minute and 1 minute.

These two datasets are referred to as the 'pre-trial' datasets in this report.

However, challenges were identified with pre-trial datasets, specifically:

- Absence of a suitable dataset for batteries, and other heat pump technologies.
- Limitations of the Electric Nation and FREEDOM datasets for the purposes of *Aggregated Datasets*: These datasets both include some limitations which introduces uncertainty in the Completeness and Resolution analysis; this is discussed in further detail in APPENDIX 1: Data Pre-Processing.

As a result, we have additionally considered how data from the Sustain-H trial itself might be used. The Sustain-H trial participant forms received were greater than anticipated, with most participants stating their ability to share asset meter and/or smart meter data to facilitate trial learnings. As a result, we have amended the requirements of our Sustain-H Data Proposal pack to request additional data from Sustain-H participants, beyond that which is strictly required to deliver the Sustain-H service. As Sustain-H is a very low value service, we have sought to make these additional requirements as palatable as possible to participants by placing the pre-processing work on Everoze; whilst this entails additional work under *Aggregated Datasets,* it has the highest chance of getting more out of participants. At the time of writing, we do not yet have responses from participants on what additional data might be shared.

As the Sustain-H data will only be available from 2021 onwards, this entails a two-stage approach to analysis. Specifically, we will run the analysis with pre-trial datasets in Q3 2020, and then append to this analysis Sustain-H data in 2021 (which, for clarity, may only involve running a subset of the full analysis presented here). An implication of this approach is that a Change Request will need to be submitted, as Sustain-H data will not be available early enough to deliver to the original timescale.

4. ANALYSIS APPROACH

4.1 PRINCIPLES

In framing our methodology, we adopted five overall guiding principles. These are as follows:

1. **Technology agnosticism:** Develop a technology agnostic method or approach, which can be applied to range of technologies and datasets collected pre-trial and from the Sustain-H trial.[1]
2. **Data application:** Make appropriate use of pre-trial datasets.
3. **Usefulness:** Ensure that data output is in a format which can inform the DDF, rather than being unduly theoretical or academic.
4. **Pragmatism:** Retain appropriate balance between simplicity and robustness in method where possible.
5. **Robustness:** Demonstrate a level of robustness of results from analysis of the pre-trial dataset, by updating the findings through Sustain-H trial data analysis.

We then defined a set of data analysis tasks that shall be undertaken. Each task is linked with a defined output. We structured these tasks according to the three DDF parameters of Resolution, Completeness and Accuracy. The task outlined in this section are priority activities to be undertaken within *Aggregated Datasets*. Additional possible tasks that were discarded are documented in APPENDIX 2: Follow-on work.

4.2 RESOLUTION

TABLE 1 lists the tasks to derive Resolution Uncertainty Factors (RUFs). The outputs will be used to inform the DDF, which is described in Section 5.

TASK	OUTPUT
1. Determine the temporal variance of the data at minutely intervals and when averaged into half hourly intervals for a portfolio.	A look up table of the mean (peak-1-minute/30-minute mean) ratio for each HH settlement period for the period considered in the analysis. The table will capture the maximum, minimum, std and mean values.
2. Quantify the impact portfolio size has on uncertainty on half hourly averaging periods.	Same output from steps 1 but for different portfolio sizes.
3. Estimate the value of RUFs for each dataset, portfolio size and resolution	Table of RUFs for each scenario, calculated using the output from Task 1 and 2.

TABLE 1: RESOLUTION METHODOLOGY TASKS

[1] The methodology will be technology agnostic. The outputs from the analysis, regardless of technology, will be assumed technology agnostic until a substantial quantity of data is available for each type of technology type (EV, ASHP and BESS). This is so that a statistically valid conclusion on technology-specific attributes can be formulated. Whilst it may be preferable to provide a table of technology-specific DQFs, it is likely that there will be insufficient data to support it. Everoze shall review this position following analysis of Sustain-H trial data.

4.3 COMPLETENESS

TABLE 2 lists the tasks to derive Completeness Uncertainty Factors (CUFs). The outputs will be used to inform the DDF, which is described in Section 5.

TASK	OUTPUT
1. Quantify the 'completeness' of each dataset for the different portfolios considered	A probabilistic estimate of the percentage completeness for each dataset under analysis
2. For a fixed portfolio size, determine the impact 98, 95, 90 and 80 % portfolio data completeness has on measured demand uncertainty.	Demand uncertainty for different percentage completeness levels, calculated using a Monte-Carlo analysis. The mean, max, median and std for each 30 minute interval is captured.
3. Estimate values of "CUFs" for each dataset	Table of CUFs for each scenario, calculated using the output from Task 1 and 2.

TABLE 2: COMPLETENESS METHODOLOGY TASKS

4.4 ACCURACY

TABLE 3 lists the tasks to derive the Accuracy Uncertainty Factor (AUF). The outputs will be used to inform the DDF, which is described in Section 5.

TASK	OUTPUT
1. Quantify the uncertainty introduced to the metered data attributable to the metering accuracy of half hourly smart meters (SMETS1 and SMETS2 meters) and the asset level meters used by participants for the trial.	Metering accuracy of smart meters (e.g. from the standards and regulatory documents) and a range of metering accuracy of asset level meters (e.g. from data sheets provided by trial participants).
2. Calculate the standard deviation of the accuracy measurements	A standard deviation calculation based on the available accuracy measurements.
3. Estimate the values of "AUFs"	Table of AUFs calculated using the output from Task 1 and 2.

TABLE 3: ACCURACY METHODOLOGY TASKS

5. METHODOLOGY

This section outlines the proposed methodology for analysing data, both from pre-existing datasets as presented in [1], and from Sustain-H trial datasets.

This section includes the following:

- Summary of Method;
- Method for analysing impact of Resolution on Quality;
- Method for analysing impact of Completeness on Quality;
- Method for analysing impact of Accuracy on Quality; and
- Assumptions and Study Limitations.

5.1 SUMMARY OF METHOD

The Aggregated Datasets sub-workstream aims to quantify the impact that varying magnitudes of measurement uncertainty have on the measured aggregated portfolio demand. It is hypothesised that this impact will be inversely proportional to portfolio size, or, in other words, that the more assets within the aggregation the more the various sources of uncertainty will tend to cancel out or diminish in terms of their proportional impact.

The method proposed is based on the following logical sequence:

1. Any 30-minute mean (average) demand value, where completeness is unknown (and unknowable) and where demand is expected to vary over time (and hence within the 30-minute period), can be improved by an understanding of:
    a. Resolution – where correcting from 30-minute resolution (for example) to determine the peak 1-minute resolution demand during that period would increase the perceived demand estimate.
    b. Completeness – where correcting for "incompleteness" would increase the perceived 30-minute average demand; and
    c. Measurement accuracy – where measurements of individual assets or smart meters are independent of each other, and normally distributed.
2. Because these factors are independent of each other, their respective impacts can be treated as such. This means they can be assessed independently, and combined using well-established statistical model(s) or simple maths.
3. Therefore, the methodology proposed will output a model of independent and 'variable' (therefore probabilistic in nature) Uncertainty Factors (UFs) for: Resolution, Completeness and Accuracy.
4. The Completeness and Resolution UFs will consist of a 'multiplier' of the 30 minute average, and the variability shall be represented by the standard deviation of the multiplier (from the available data).
5. For Accuracy, the multiplier will be set as 1. However, there will be a standard deviation which shall be defined by accuracy measurement based on available information (e.g. from datasheets and Sustain-H trial participant data, and the number of measurements made).
6. The UFs will be combined into a single DQF estimate (as represented in **Error! Reference source not found.**) for each combination of elements considered, with outputs in a tabulated form:
    a. Dataset (for the pre-trial datasets: Electric Nation and FREEDOM and also Sustain-H data);
    b. Portfolio size;
    c. Completeness; and
    d. Resolution.
7. The DQFs will be used to establish DDFs (commercial impact), although this will be undertaken by WPD outside of this sub-workstream.

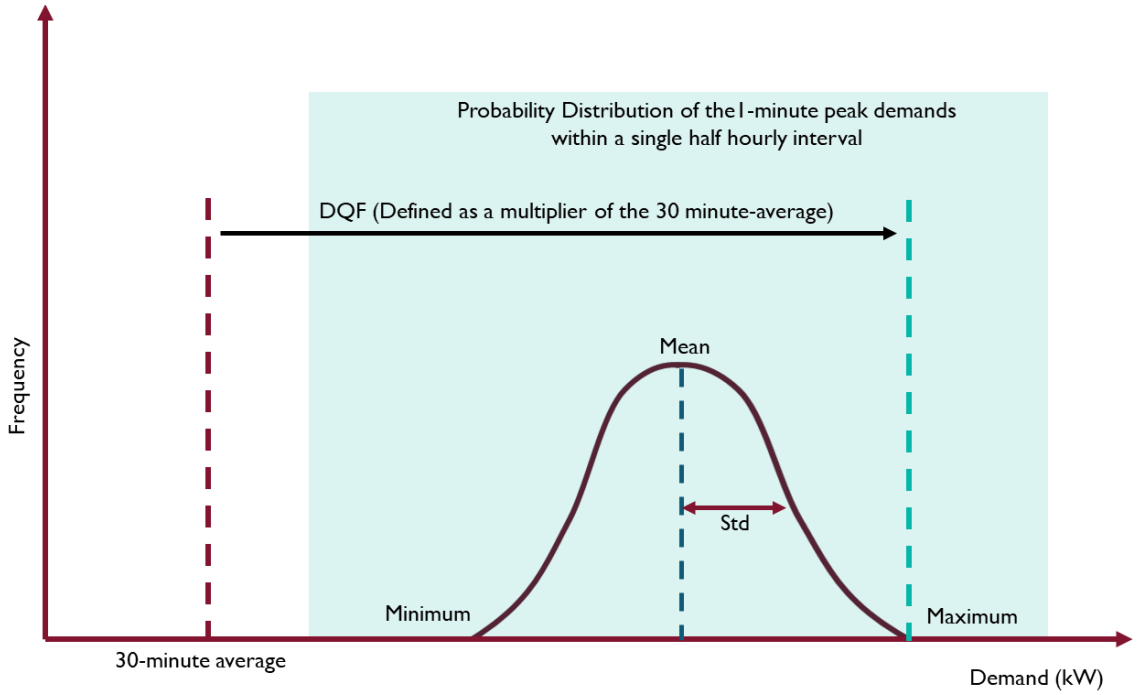


FIGURE 4: RELATIONSHIP BETWEEN THE 30-MINUTE AVERAGE DEMAND AND DISTRIBUTION OF 1-MINUTE PEAK DEMANDS, WHERE ALL DATA IS FOR A GIVEN HALF HOURLY PERIOD EACH DAY.

5.2 SUMMARY OF DELIVERABLES

Two reports will be delivered: 1) following the pre-trial dataset analysis and 2) following Sustain-H data analysis, which contain the following:

- Detailed description of the process applied, including the steps taken to process the data;
- Tables of DQF;
- Charts/figures etc from the analysis.

The intent is to provide sufficient information for the analysis process applied to be reproduceable and adaptable for analysis of other datasets by WPD. Converting the analysis code into a tool which is client-ready would require extra time and is not within the scope of this workstream; however, it is possible to provide such a tool with additional budget.

5.3 RESOLUTION

There are four key parts to the analysis:

- Determining variance of the data at 1-minutely and for 30-minutely average data;
- Quantifying the impact portfolio size has on uncertainty for the different averaging periods; and
- Estimating "RUFs" for the derivation of Data Quality Factor (DQF). These are described below.

It is noted that, whilst all items shall be undertaken on the pre-trial datasets, it is the intention that only the first item shall be undertaken on the Sustain-H dataset (for a comparable portfolio size) to test against RUFs determined from the pre-trial datasets.

5.3.1 Determine the variance of the data at 1-minutely intervals and when averaged into half hourly intervals for a portfolio.

It is assumed that the high-resolution dataset is of at least 30-minute frequency. It is known that, following pre-processing, the Electric Nation data can be analysed at a 1-minute resolution (with some pre-processing assumptions, as described in APPENDIX 1: Data Pre-Processing), and it is expected that FREEDOM data shall be made available at 1-minute resolution, provided by PassivSystems.

It is currently unclear whether, or how much, Sustain-H trial data may be available on a minutely basis. Depending on data provision, the RUF could be updated using Sustain-H trial data using this approach.

Analysis steps are presented below:

Step 1: For a fixed population size within a defined period (maximising the size and period of the dataset as appropriate), sum the minutely portfolio electricity demand. The output will be the net portfolio demand (in kW) for each minute over the sampling period (e.g. for one month, assuming 1 minutely resolution, the number of data points will be up to 31 days * 24 hrs * 60 mins = 44,640 data points). Dependent on coverage, we may need to "correct" for variation in participants, although we will seek to ensure consistency and completeness within the fixed population.

Step 2: Everoze aims to quantify how the ratio of peak-1-minute / 30-minute mean varies for each HH settlement period (48 in total), and also for each Delivery Period for Sustain-H. The following steps will be taken to analyse the (peak-1-minute/30-minute mean) ratio for each HH settlement period for the period considered in the analysis:

a. Calculate the mean (30-min) demand for each HH period within the time frame considered[2].
b. For each 30-min interval across the timeframe considered, calculate the (1-min peak / 30-min mean) ratio.
c. For each half hourly settlement period (48 in total), derive the probability distribution of the (peak-1-minute/30-minute mean) ratios using the output from b, to produce 48 probability distributions.
d. For each of the 48 settlement periods, derive the following values from the probability distributions and write the values in a look up table (for an example, see Table 4) :
    1. Mean (peak-1-minute/30-minute mean) ratio;
    2. Maximum;
    3. Minimum; and
    4. Standard Deviation.
    5. P95 exceedance cases.

Investigating P95 or P90 exceedance cases will be used to support or validate the use of Standard Deviation, and this shall be considered during the analysis.

Step 3: Everoze shall extract the "multiplier" (in this case the Mean (peak-1-minute/30-minute mean) ratio) and corresponding maxima, minima and standard deviations for each of the 48 half hourly intervals**.** The Mean ratio is expected to be a value greater than 1, because the 1-minute peak within a 30-minute time period will be greater than the average demand during the respective 30-minute period.

If data completeness during the analysis period is deemed to have a material impact on the analysis, appropriate correction factors may need to applied. This will be considered during the analysis.

	Half hourly interval								
Peak (1 min) / mean (30 min)	00:00–00:30	00:30 – 1:00	01:00-01:30	01:30-2:00	2:00-2:30	…	22:30-22:00	23:00-23:30	23:30-00:00
Mean									
Maximum									
Minimum									
Standard Deviation									

TABLE 4: OUTPUT TABLE RESOLUTION

[2]Implicit assumption: for fixed population size, the variance in completeness of the dataset is small so that there is nedligible noise within the results. Using a dataset close to 100% completeness will improve the validity of this assumption.

5.3.2 Quantify the impact portfolio size has on half hourly averaging uncertainty

Step 1: Randomly select a sub-group of the data population from those analysed in Section 5.3.1 to create reduced population sizes (for example, reducing a dataset of 100 participants to 25 and 50 for further analysis).

Step 2: Repeat Steps 1 to 3 from Section 5.3.1 on the different population sizes to assess how the results vary.

5.3.3 Estimate value of "RUFs"

Through analysis of the findings in the analysis described above, estimate the value of "RUFs" for each:

1. Dataset
2. Portfolio size; and
3. Resolution.

5.4 COMPLETENESS

There are three key parts to the analysis:

- Quantify the 'completeness' of the datasets from different portfolios considered;
- Determine the impact of different levels of data completeness on the uncertainty of measured portfolio demand representing the 'real' portfolio demand; and
- Estimating "CUFs" for the Data Quality Factor equation.

These are each described below.

5.4.1 Quantify the 'completeness' of the dataset for the different portfolios considered

Step 1: For a fixed period (i.e. 1 month), quantify the percentage completeness for each half hourly period across the portfolio as a single metric for the dataset under analysis.

Step 2: Plot the probability distribution of the percentage completeness across the fixed period.

Step 3: Fit a simple mathematical distribution to the probability distribution (e.g. a Weibull curve).

5.4.2 For a fixed portfolio size, determine the impact 98, 95, 90 and 80 % portfolio data completeness has on measured demand uncertainty

Potential invalid datapoints in a HH metered dataset (for example) may take the following forms:

1. An obviously invalid value (e.g. outside conceivable range)
2. Null value
3. Zero value (could or could not be valid)
4. A realistic value which may not capture the full duration of that time period (e.g. only 15 minutes of data in a 30 minute period)

In order to estimate the impact of Completeness on uncertainty, it is necessary to artificially introduce "incompleteness" into a complete dataset and quantify the difference in measured demand for varying levels of completeness. This requires the starting dataset to be 100% complete before introducing null (or back filled) values randomly.

When aggregated over a portfolio, this could have the effect of artificially lowering the portfolio metered value for the affected periods, where the 'real' portfolio demand would be higher than this metered value.

Steps 2 to 6 set out steps to quantify the percentage uncertainty for different completeness and population size scenarios.

Steps 2 to 6 should first introduce incompleteness as null values. The incomplete data points can then be back filled by the mean of the valid data points.

It is assumed that the data received from trial participants will be of 30 minute resolution, although other resolutions can be applied.

Step 1: Set up the starting dataset to be 100 % complete so that random incompleteness can be introduced artificially.

Step 2: For a fixed population size (e.g. 100 homes), sum the 30 minute average electricity demand for each 30 minute interval across the timeframe considered. The output would be the total demand for each 30 minute interval.

Step 3: Randomly introduce an average completeness (e.g. 98%) into the dataset by allocating 2% of electricity demand readings as null [0 kW] values by using a skewed-normal (Weibull) distribution which has been derived from, and checked against, appropriate data (as described in Section 5.4.1). Sum the electricity demand for each 30 minute interval.

Step 4: For each 30 minute interval, divide the output from step 2 with the output from step 3 to give a demand difference for each 30 minute period (The numerator is always 100% complete 30-minute demand, therefore by dividing the 100 % complete 30 minute demand by the 98 % complete 30-minute demand, the ratio for each 30 minute interval should be greater than 1, or equal to 1, depending on the random allocation of incompleteness in the data).

Step 5: Repeat Steps 3 to 4 for several iterations (e.g. 100 iterations) for different cases of random introduction of incompleteness, in the form of a Monte Carlo analysis.

Step 6: For each 30 minute interval, create a probability distribution from the full output of Step 5 and calculate the following (TABLE 3 captures the outputs as a table for one 30 minute interval):

1. Mean;
2. Maximum;
3. Minimum; and
4. Standard Deviation.

% Completeness	Std	Maximum	Minimum	Mean
98	Average demand 100 % C / average demand 98 % C	…		
95	…			
90				
80				

TABLE 3: OUTPUT TABLE QUANTIFYING DELTA DEMAND FOR A 30 MINUTE INTERVAL

For the pre-trial datasets, we shall consider whether it is necessary to undertake this with different participants (portfolio sizes), and potentially even for differing data resolutions in order to provide sufficient information to support the derivation of the DQFs as discussed in Section 5.1.

5.4.3 Estimate values of "CUFs"

Through analysis of the findings in the analysis described above, estimate the "CUFs" for each dataset.

5.5 ACCURACY

There are three key parts to the analysis:

- Quantify the 'Accuracy' of different meter types;
- Calculate the standard deviation of this uncertainty; and
- Estimating "AUFs" for the Data Quality Factor equation.

These are each described below.

5.5.1 Quantify the uncertainty introduced to the metered data attributable to the metering accuracy of half hourly smart meters (SMETS1 and SMETS2 meters) and the asset level meters used by participants for the trial.

Everoze shall review metering standards for other ancillary services such as FFR and determine appropriate parameters for metering accuracy.

Eveorze shall discuss with WPD and trial participants, where possible, to ascertain any information as to typical metering accuracy.

In the event that there are varying estimates made available, Everoze shall undertake a broad sensitivity analysis to examine the consequences of varying meter accuracy on overall Data Quality for a varying portfolio size.

5.5.2 Calculate the standard deviation of the accuracy measurements

The standard deviation of the combined accuracy measurements will be assessed based on available information.

5.5.3 Estimate the values of "AUFs"

Through analysis of the findings in the assessment described above, define the range of "AUFs" for a range of meter options.

5.6 ASSUMPTIONS AND STUDY LIMITATIONS

5.6.1 Technology-specific Data De-rating Factor

Any conclusions that Everoze can make on the variation per technology are likely to have low validity due to the small range of technologies assessed, and their respective contexts. In order to determine a technology-specific Data De-rating Factor, the population sizes and a variety of suppliers/providers within each technology must be substantial in order to make evidence based and non-bias, technology-specific conclusions. The Sustain-H trial participants are predominantly entering with smart EV chargers, a few hundred batteries and fewer than 10 heat pumps. Consequently, conclusions which feed into the Data De-Rating factor will be technology agnostic at this stage. However, this outcome helps to formulate the next steps.

5.6.2 Sustain-H Trial Data

There is a risk that Sustain-H participants will not provide individual meter data for the benefit of the Sustain-H learnings and consequently Aggregated Datasets sub-workstream. This is because data provision to support the Aggregated Datasets sub-workstream is optional. The risk of receiving no data is low, because all Sustain-H Participants have informed Everoze of their interest in providing either or both Smart Meter of Asset level meter data.

A second risk is that the Quality of the Sustain-H data will be poor (e.g. a high degree of erroneous values). After the data is received from each participants, the data will be checked for obvious inconsistencies and shortcomings.

5.6.3 Pre-Trial Datasets

Refer to APPENDIX 1: Data Pre-Processing for a list of assumptions and limitations regarding the pre-trial datasets.

6. NEXT STEPS

The proposed timelines for remaining scope items for the Aggregated Datasets sub-workstream are outlined in the table below.

SCOPE ITEM	DELIVERED BY
1. Conduct Analysis outlined in Section 5 of this Report.	30/09/2021
2. Form recommendations	31/10/2021

This represents an extension of the original timeline to accommodate receipt of Sustain-H data, and assumes that a Change Request can be submitted to sign off on the timeline change. The timescale is dependent on two factors:

- **Time and quantity of data received from the Sustain-H trial participants**. It is not yet known with certainty how many Sustain-H participants will be willing to share additional data to support the Aggregated Datasets workstream. The data could come in on a rolling monthly basis, be intermittent, or none could be received. Everoze is maximizing chances of receiving appropriate data through scheduling one-to-one calls with participants on data.
- **Sustain-H trial end date**: 31/07/2021.

As it currently stands, the timescale of interim deliverables for scope item 1 are as follows:

INTERIM DELIVERABLES FOR SCOPE ITEM 1	DELIVERED BY
Conduct analysis on the two pre-trial datasets: Electric Nation and FREEDOM	31/10/2020
Collect Sustain-H trial participant data on a monthly basis throughout the trial duration.	31/07/2021
Pre-process the first datasets received from each participant the month following retrieval. This is to allow time for assessment of the data format so any subsequent data requests could be made, if necessary, before any further data is received.	31/07/2021
Run analysis on received trial datasets, as outlined in Section 5 of this Report.	31/08/2021
Deliver Scope item 1	31/10/2021

7. REFERENCES

1. Everoze, WESTERN002-AD-R-01-B, "FutureFlex: Supplier Trial – Data Review", Revision B, 29th May 2020
2. Western Power Distribution, Electric Nation, "Electric Nation Customer Trial - CrowdCharge Transaction Data", 24th October 2019
3. PassivSystems, FREEDOM, "FREEDOM Trial Data", 18th April 2018

8. APPENDIX 1: Data Pre-Processing

8.1 PRE-TRIAL DATASET ANALYSIS SUMMARY

HEADLINE FINDINGS

To verify the assumptions from [1], additional analysis of the pre-trial datasets was carried out. The goal was to ensure that the data can be pre-processed into an appropriate format, as intended by the methodology.

Findings:

- FREEDOM and Electric Nation (CrowdCharge) datasets can be used to inform the DDF, but there are inherent assumptions attributable to the data quality.
- The DDF outputs should be updated using Sustain-H trial asset level and smart metering data.
- No additional third-party datasets were identified since deliverable [1].

The full details are captured below.

Two pre-trial datasets are suitable for usage. These datasets are:

- FREEDOM: 1, 5 and 30 minutely hybrid gas boiler-ASHP data (continued dialogue with PassivSystems indicates more data is available from the FREEDOM project).
- Electric Nation: includes EV charger start and stop times at minute resolution.

The pre-trial datasets have limitations, which drive a need for validation with Sustain-H trial data. The pre-trial datasets will help shape the analytical model and provide valuable insights to inform the DDF. However, the datasets have inherent limitations and therefore outputs needed to be updated using Sustain-H trial data. The limitations are attributed to:

- Trial interventions: correcting for the interventions introduces additional demand uncertainty;
- Null data: Section 8.2.2 lists the data removed during Electric Nation pre-processing;
- Coarse data reading frequency; and
- Data inapplicability to the Sustain-H trial. The FREEDOM dataset is a hybrid gas boiler-ASHP system. None of the Sustain-H trial participants have entered in hybrid ASHP systems, despite the technology being eligible to participate. However, because the methodology is technology agnostic, this should not be an issue.

8.2 ELECTRIC NATION DATA

Pre-analysis was carried to inform the usefulness of the Electric Nation dataset for the purposes of the Aggregated Datasets workstream. The steps carried out to remove null data prior to analysis and results are captured in Section 8.2.2 below.

8.2.1 Assumptions

Everoze made a number of assumptions on the Electric Nation dataset:

1. The electricity demand for each transaction is constant during the charging period. Electric Nation data includes the start and end charging times to the nearest minute for each transaction (i.e. charging event), rather than the minutely demand. Therefore Everoze will assume that the EV charging demand for each transaction is constant between timestamps. However, this is highly unlikely to be the case because for example, when the car battery reaches 80 % charging capacity, the rate of charge decreases.
2. The 'behaviour' of charging illustrated from Electric Nation data is representative of the wider population. However there are caveats to this assumption because:

a. There are interventions influencing the charging rate which are not captured from the averaged minutely demand estimations, such as limiting the current supplied to chargers to relieve network constraint; and
b. The charging rate is unlikely to be constant and therefore the peak demand captured from Electric Nation data might be higher or lower than the actual peak demand. This means resolution uncertainty calculated using Electric Nation data might be higher or lower than proven by the trial calculations.

3. The conditions within the datasets capture the full variation of usage profiles required to fully assess the DQFs and the DDFs respectively.

8.2.2 Pre-Processing

Electric Nation Data Format Summary

Each 'transaction' i.e. charging event is provided as a single row in the Electric Nation data file. The columns provide information on the start charging (ActiveCharging_Start) and end charging (EndCharge) time, and the total electrical demand during the charging event (Consumed kWh). Therefore the energy consumed during the charging period is a minutely average. Ideally, the data would be minutely 'actual' power consumption rather than an average demand during charging period.

Electric Nation Data Pre-Processing

Data pre-processing was undertaken in 5 stages:

1. Removed data points whereby the consumed kWh was greater than 100 kWh. This is a result of inaccuracies in the data. The largest battery capacity participating in the project was 100 kWh, therefore any charging events greater than this were an anomaly.
    a. Events where the ConsumedkWh < 0.5 were kept in despite the total charge being marginal, because this indicated that the car was plugged in and available for DSO services even if the car was charged.
2. Removed any data where Participant ID = Null. In these instances, the participant could not be linked to a charger which is required to quantify the total number of participants charging on any one date.
3. Removed instances where ActiveCharging_Start = Null. This was caused due to internet connectivity issues.
4. Removed instances when EndCharge = Null. This was always the case if ActiveCharging_Start = Null because the EndCharge time could not be approximated. Additionally this was caused if the internet was down during the end of the charging period.
5. Checked CrowdCharge note (incorrectly calculated EndCharge). CrowdCharge noted that there are instances when the EndCharge time was not calculated correctly in the datasheet. EndCharge was estimated (by CrowdCharge analysts) as the time when charging ended based on the meter value data (looking for sustained reductions in the current being drawn in the absence of demand management). Some of these calculations were inaccurate because the total consumed kWh over the charging duration was greater than technically possible given the charger capacity. For example, for a single transaction using a 7kW charger, if ActiveCharging_Start was 17:00 and EndCharge was 18:30, but 21kWh of energy was transferred then it's likely that EndCharge is inaccurate (because it isn't possible to add 21kWh in 1.5 hours using a 7kW charger).
    a. Everoze performed the following to check if the 'Charge time' was too short:
        i. ChargeTime = EndCharge – ActiveCharging_Start
        ii. ChargingRate = ConsumedkWh/ChargeTime
        iii. If ChargingRate>Battery Rating, then the end charging time is incorrect.
        iv. There are 379 transactions where this was the case, which is 0.53% of the data.

    To overcome this error, the EndCharge time was re-calculated for the rows where the above formula was true by the following:
        v. NewChargeDuration = ConsumedkWh/CarkW(rating of the participant's vehicle, either 3.6 kW of 7 kW which is provided per row).
        vi. NewEndCharge = ActiveChargingStart + NewChargeDuration

    Assumption: that the rate of charge was constant at 7kWh or 3.6 kWh throughout the charging period.

Table 5 captures the number (and percentage) of transactions removed from the input Electric Nation data file during pre-processing stages 1 to 4. Stage 5 is not included because data was not removed from the input file, instead the EndCharge times were corrected and replaced.

Stage	Transactions Removed from Input File	Cumulative % Removed
1	21	0.03
2	3899	5.47
3	19707	27.65
4	20855	29.26

TABLE 5: ELECTRIC NATION DATA PRE-PROCESSING: TRANSACTIONS REMOVED DURING EACH STAGE LISTED IN SECTION 8.2.2.

8.2.3 Results from pre-processing

Factors which influenced the decision to use the Electric Nation (CrowdCharge) dataset included:

1. 70.74 % of the initial dataset (71,264 transaction) remained after correcting for the null data, meaning the data remaining was still substantive.
2. There were three trials conducted as part of the Electric Nation project. Trial 1 featured minimal interventions, meaning that the 'trial' data is close to a 'real-world' scenario and therefore the impact of trial interventions on the UF calculations are low if using Trial 1 data.
3. The Completeness and Resolution UF calculations, written in Section 5, can be performed on the data remaining. The results might be slightly inaccurate because the trial data is not 'real-world', however, the UFs (and subsequently DQF) will be updated following retrieval of Sustain-H trial datasets, and the data is of sufficiently high quality to provide a good level of certainty on the UFs despite the assumptions.

8.3 FREEDOM DATA

Pre-analysis was carried to inform the usefulness of the FREEDOM dataset for the purposes of the Aggregated Datasets sub-workstream.

Everoze reviewed the FREEDOM datasets (30-minute and 5-minute resolution datasets and a sample 1-minute dataset) via interpretation of the dataset and also by reviewing the FREEDOM datasheet (which described the data within the dataset). The result of this analysis provided relatively substantive confidence on the data capabilities. The more in depth pre-analysis of the Electric Nation data provided a good understanding of 1-minute resolution data capabilities (regardless of technology). Minutely data is expected to be received from PassivSystems in the near future. The pre-analysis led to a series of caveats in the data being identified, which led the following assumptions being deduced.

8.3.1 Assumptions

Everoze made a number of assumptions on the FREEDOM dataset:

1. The trial participant behaviour (heat set point) is representative of a wider population. However, there are caveats to this assumption because the trial introduced a series of interventions which influenced the optimisation. The hybrid gas boiler-ASHP systems were set up to optimise the heat output for least cost to

customer whilst meeting the heating set point. Therefore the data might not reflect 'real world' data/behaviour. Correcting for the interventions would induce errors which would be difficult to quantify and so Everoze has decided not to correct for this, which might induce significant caveats in the UF calculations.

2. The conditions during the trial capture the full variation required to fully assess the DQFs and the DDFs respectively.
3. The hybrid gas boiler ASHPs used in the trial (3 in total) behaviour similarly to non-hybrid ASHP systems

9. APPENDIX 2: Follow-on work

Appendix 2 introduces potential follow-on tasks which are not being completed for the Aggregated Datasets analysis, and therefore did not make it into Section 4 of this report. The tasks were marked as lower priority due to:

1. Marginal value to effort.
2. Data available means some tasks are unlikely to be answered with high outcome certainty.
3. Adding more complexity than necessary at this stage.

It is recommended that these tasks are researched upon conclusion of the Aggregated Datasets workstream, as follow-on work. Within this section, an additional factor is introduced which may influence the Data De-Rating factor: Half Hourly Household Visibility.

Uncertainty Factor	Task Included Within Methodology	Possible Tasks For Follow-On Work
Completeness	For a fixed portfolio size, determine the impact 98, 95, 90 and 90 % portfolio data completeness has on measured demand uncertainty (using 30 minute resolution data)	1. Quantify impact of different averaging periods (5, 10 and 15 minute) 2. Quantify (1) for i) EVs ii) heat pumps and iii) batteries. 3. With different sample datasets, see whether there is a variation in the output from (1) and see whether this is a feature of the technology and metering option, or tied to a particular operator's approach and not fundamental to the technology/metering option.
Resolution	Determine the variance of the data at 1-minutely intervals and when averaged into half hourly intervals for a portfolio.	1. Determine the temporal variance of the data when averaged into 5, 10 and 15 minute periods (in addition to 1 minute and half hourly) 2. Quantify the uncertainty from the output from (1) for the different technologies considered (using available datasets): i) EVs, ii) heatpumps, and iii) batteries. 3. Measure any technology specific attributes
Accuracy	Quantify the uncertainty introduced to the metered data attributable to the metering accuracy of half hourly smart meters (SMETS1 and SMETS2 meters) and the asset level meters used by participants for the trial.	1. Analyse how diverse the asset meter accuracy measurements are on a per technology basis (see whether the variance (if any) can be attributed to the technology itself, or other variables
Half-Hourly Household Visibility	N/A	Perform analytics on the portfolio consumption behaviour to see whether households show a general increase/decrease in their overall consumption behaviour (ie., non-flex loads) when the flexibility assets are dispatched. Explore the answer: 1. Relative to periods when the flexibility assets are not operational. 2. Focus on the evening peak demand period. 3. Research whether consumers are increasing their household demand (say turn the washing machine on, etc) for periods when say the batteries are discharging.

TABLE 6: POTENTIAL FOLLOW-ON WORK.