Smart meter consumption data: Technical documentation

Creation date	2021-03-29
Edition	02
Author	Ellen Webborn
Project	Smart Energy Research Lab (SERL)
Organisation	University College London (UCL)

Table of Contents

Introduction

This document describes the half-hourly and daily datasets available to researchers with secure access to the SERL Observatory datasets, along with two data quality summary tables. The data were collected from the earliest date available. The datasets described in this document are:

- *serl_smart_meter_daily_edition{number}.csv*: daily electricity and gas readings with some additional derived columns
- *serl_smart_meter_hh_edition{number}.csv*: half-hourly electricity and gas readings with some additional derived columns (note that this dataset also includes reactive readings and export readings where available)
- *serl_smart_meter_rt_summary_edition{number}.csv*: data quality summary for each read type for each participant (such as number of errors found by type) and basic read statistics (such as mean and maximum)
- *serl_participant_summary_edition{number}.csv*: data quality summary for each participant (less detail than the read-type summary data, but also including basic participant information such as region and number of questions answered on the survey)

where *{number}* is the number of the data release, *e.g.* “02” (note that the first data release files are labelled with the release date rather than edition number).

This document is structured as follows: we start with some basic information about how the data were collected and the different types of reading available, then we describe the two smart meter data tables (daily and half-hourly). Next we define the different types of error flag created, and finally we describe the data quality summary tables (at the read-type level and the participant level).

Data collection

Half-hourly and daily smart meter readings are stored on the smart meter, and accessed by the Smart Energy Research Lab (SERL) as follows. The University of Essex (UK Data Archive) uses a DCC adaptor service provided by CGI to communicate with the DCC, which acts as a pipe to communicate the smart meter readings to CGI, who send the readings to the UK Data Archive. This happens every night to collect data from the previous day.Not all properties have a gas meter we are able to access - for example, if there is only a SMETS2 electricity meter but not a SMETS2 gas meter, or if the property does not have mains gas. Check the EPC and survey data to identify properties which have gas central heating but no SERL gas data if this may affect your analysis.

Smart meter read types

Smart meter read types are defined by two variables: 'deviceType' and 'readType'. Together they combine to define the type of smart meter data. The full list of smart meter data types are shown in the table below. Note that 'GPF' stands for 'Gas Proxy Function' (a proxy for the gas meter) and 'ESME' stands for 'Electricity Smart Metering Equipment' (the electricity meter).

Table 1: Smart meter data types, defined by 'deviceType' and 'readType'.

deviceType	*readType*	*Units*	*Description*
ESME	DL	Wh	Daily active electricity import
GPF	DL	m^3	Daily gas import
ESME	AI	Wh	Half-hourly active electricity import
ESME	RI	Wh	Half-hourly reactive electricity import
GPF	AI	m^3	Half-hourly gas import
ESME	AE	Wh	Half-hourly active electricity export
ESME	RE	Wh	Half-hourly reactive electricity export

Changes since the previous edition

In addition to the inclusion of participants from wave 2 recruitment and the extension of data to 31st October 2020, a number of changes have been made to the processing of the raw data. The main changes are:

- Previously we believed daily reads to be taken at midnight UTC time, whereas our analysis has shown that the reads are taken at midnight local time (changes during British Summer Time (BST)). Half-hourly reads are taken in UTC, therefore half-hourly sums are now converted to local time to compare with daily reads.
- The error flags in the first release reflected the validity of both the read and the read time. This can be ambiguous, so the error flags now relate only to our beliefs about the validity of the read (value). The *Valid_read_time* flag indicates the validity of the read time.

- Error flag value 3 has been added for instances where one type of read is recording at the wrong time and another type of read (being flagged) is 'missing' because the time is invalid so no read is required. This 'missing' read should be ignored.
- Some meters were found to be recording in Wh (the correct units) and after a meter replacement/upgrade the units changed to kWh. This was not captured in the previous data release.
- To cut down on the size of the datasets some columns previously added have been removed where considered superfluous.
- Where no reads exist at a given time there will be no row in the dataset for that participant. In the latest release we now create a row in the daily data if there is a valid half-hourly sum for the day, so that researchers can use this sum more easily to impute the missing daily value.
- The half-hourly sum columns only contain values if the half-hourly sum is deemed to be valid (the correct number of half-hourly reads all valid). Otherwise the value is NA (not the case previously).
- Limits used to class a read as 'very high' have been refined (see section below).
- We provide a file *bst_dates_to_2024.csv* detailing the start and end dates of British Summer Time to help researchers identify potentially anomalous dates in the datasets and consider the impact of clock changes on their research.

We continue to analyse and improve the SERL datasets and we aim to strike a balance between improving the data and minimising inconvenience to researcher caused by changes. We welcome feedback to improve our processes.

Smart meter data tables

Half-hourly data

The half-hourly data table has 16 columns. The fields are described in the table below. While the original data have not been modified, additional columns have been added to flag potential errors (see the Error Flags section below) and convert between units. This processing was done using R version 4.0.1 (2020-06-06). Code used for processing will be made available shortly on the SERL github repository github.com/smartEnergyResearchLab.

Note that if no data were returned for any meter at a particular time then that row will be missing from the dataset rather than being an empty row. The 'class' field is the R class (e.g. R calls a Boolean a logical).

Table 2: Half-hourly data fields. Error flags are defined in the Error Flags section below.

Field	*Description*	*Units*	*Class*	*Example value*	*Variable type*
PUPRN	Pseudonymised participant identifier	NA	character	1VUXXXF1	Assigned
Read_date_effective_local	Date of read (same as date of Read_date_time_local unless read taken at midnight, then the previous day since data pertains to the previous day)	%Y-%m-%d	Date	2019-11-01	Derived
Read_date_time_local	Time read taken (local time: GMT or BST)	%Y-%m-%d %H:%M:%S tz	character	2020-07-02 00:03:30 BST	Derived
Read_date_time_UTC	Time read taken in UTC	%Y-%m-%d %H:%M:%S	POSIXct, POSIXt	2020-07-02 00:02:30	Primary
Valid_read_time	FALSE if read time is not on the hour or half hour, otherwise TRUE	NA	logical	TRUE	Derived
Elec_act_imp_flag	Half-hourly electricity active import error flag	NA	numeric	-2	Derived
Elec_react_imp_flag	Half-hourly electricity reactive import error flag	NA	numeric	1	Derived
Elec_act_exp_flag	Half-hourly electricity active export error flag	NA	numeric	-4	Derived
Elec_react_exp_flag	Half-hourly electricity reactive export error flag	NA	numeric	2	Derived
Gas_flag	Half-hourly gas import error flag	NA	numeric	0	Derived
Elec_act_imp_hh_Wh	Half-hourly electricity active import read	Wh	integer	109	Primary
Elec_react_imp_hh_varh	Half-hourly electricity reactive import read	varh	integer	15	Primary
Elec_act_exp_hh_Wh	Half-hourly electricity active export read	Wh	integer	65	Primary
Elec_react_exp_hh_varh	Half-hourly electricity reactive export read	varh	integer	14	Primary
Gas_hh_m3	Half-hourly gas import read	m^3	numeric	0.244	Primary
Gas_hh_Wh	Half-hourly gas import read in Wh using standard conversion, assuming calorific value = 39.5	Wh	numeric	2737.835	Derived

Daily data

The daily data table has 14 columns. The fields are described in the table below. While original data has not been modified, additional columns have been added to flag potential errors (see Error Flags section below) and to convert between units. This processing was done using R version 4.0.1 (2020-06-06). Code used for processing will be made available shortly on the SERL github repository github.com/smartEnergyResearchLab.

Note that if no data were returned for any meter on a particular day then that row will be missing from the dataset rather than being empty. The exception is if there were no daily reads but the right number of half-hourly reads (48 unless the clocks changed) to sum to the daily total. In these instances the rows have been added to allow for easy imputation of a missing daily read with the sum of the half-hourly reads. For comparison between daily reads and daily sums, half-hourly data requires conversion (provided) to local time. A csv file is provided *(bst_dates_to_2024.csv)* which lists the start and end dates of British Summer Time (BST) for reference so the number of half-hours expected on each day is clear. Researchers may wish to avoid clock change days in their analysis, or take them into consideration.

The daily data includes fields added for comparison between the daily readings and half-hourly readings for the same days. As described in Table 3 there are columns for the sum of the half-hourly readings if there were 48 valid reads taken at the right times (or 46 or 50 reads if the clocks changed), the difference between the half-hourly sum and the daily read (if both exist and are valid), and a 'sum_match' column coded to state the condition of this match. The sum match codes are defined in Table 6 in the Error Flags section. Note that reads are considered invalid if taken at the wrong time (not at midnight for daily reads; not on the hour or half hour for half-hourly reads).

Table 3: Daily data fields. See the Error Flags section below for definitions of the error flags used in this table.

Field	*Description*	*Units*	*Class*	*Example value*	*Variable Type*
PUPRN	Pseudonymised participant identifier	NA	character	1VUXXXF1	Assigned
Read_date_effective_local	Date that read relates to (in local time): previous day, unless after midday (then same day)	%Y-%m-%d	Date	2019-11-01	Derived
Read_date_time_local	Time and date of read (local time). Time not stated if at midnight	%Y-%m-%d	POSIXct, POSIXt	2019-11-02	Primary
Valid_read_time	TRUE if reading was at midnight, otherwise FALSE	NA	logical	TRUE	Derived

Field	*Description*	*Units*	*Class*	*Example value*	*Variable Type*
Elec_act_imp_flag	Daily electricity active import error flag	NA	numeric	-2	Derived
Elec_sum_match	Error code for whether the sum of half-hourly electricity active import matches the daily electricity read	NA	numeric	1	Derived
Gas_flag	Daily gas import error flag	NA	numeric	2	Derived
Gas_sum_match	Error code for whether the sum of half-hourly gas import matches the daily gas read	NA	numeric	0	Derived
Elec_act_imp_d_Wh	Daily electricity active import read	Wh	integer	5839	Primary
Unit_correct_elec_act_imp_d_Wh	Daily electricity active import read corrected from kWh to Wh where kWh reporting is suspected, otherwise equals Elec_act_imp_d_Wh	Wh	integer	5839	Derived
Elec_act_imp_hh_sum_Wh	Sum of half-hourly electricity active import reads for this date (NA if there were not 48* valid reads). *46 required when the clocks go forward, 50 when the clocks go back.	Wh	integer	5742	Derived
Gas_d_m3	Daily gas read	m^3	numeric	8.214	Primary
Gas_hh_sum_m3	Sum of half-hourly gas reads for this date (NA if there were not 48* valid reads). *46 required when the clocks go forward, 50 when the clocks go back.	m^3	numeric	8.763	Derived
Gas_d_kWh	Daily gas import read in kWh using standard conversion, assuming calorific value = 39.5	kWh	numeric	92.16628	Derived

Error flags

Flags for individual reads

Table 4 shows the meaning of each error flag value. These flags are used for both daily and half-hourly reads. The data quality report gives details about the number of each error found within the data along with other descriptive statistics. Note that the first data release (2020-08) had slightly different error flags (no flag 3 and a flag for invalid read times). Importantly, the error flags are now split so that **these flags relate only to the quality of the read** (missing, too high etc.) and **not the validity of the read time**. The 'Valid_read_time' variable indicates whether the read was at the correct time (or not). Therefore to filter on valid reads at valid times, use both the relevant error flag and the Valid_read_time flag.

Table 4: Error flags and their meanings.

Flag	*Meaning*	*Details*
3	Ignore	Invalid read time and no read - row exists for a different read type so ignore
2	No meter	The gas (or very rarely electricity) meter does not exist in the DCC inventory
1	Valid	The read exists and does not meet any of the other error flag criteria, thus presumed valid (although may not have a valid read time - check separately)
0	Missing	The read should exist but is missing
-1	Max read	The read is (presumably) the largest storable number on the meter - details below
-2	Very high but not max	The read is higher than plausible but not a 'Max read' - see section below for thresholds
-3	Negative	The read is negative (none found)
-4	Elec in kWh	The electricity read was reported in kWh rather than Wh - details below

'Max reads' (error flag -1)

We discovered that multiple participants have some electricity readings as 16777215 Wh and multiple gas readings of 16777.215 m^3. These numbers are all 1s in binary which implies they are the maximum read the (32-bit) meter can store, and likely due to some technical error. We call this type of error 'Max read'. Note that in a few cases the number is the max read stored in 64 bits - these are replaced by the 32-bit maximum to save memory.

Very high reads (error flag -2)

For gas and active electricity reads we flag if the reading is larger than we deem plausible, attempting to be cautious with our definition of 'plausible'. The following table shows our definitions of what constitutes a 'very high' read (so long as the read is not high enough to be a 'max read').

Table 5: Limits used to determine if a read is high enough for a 'very high read' flag (-2).

deviceType	*readType*	*High Read Limit*	*Units*	*Assumptions*
ESME	DL	1,152,000	Wh	Max 200A fuse, 240V
ESME	AI	24,000	Wh	Max 200A fuse, 240V
ESME	AE	2,000	Wh	4kW max PV capacity
GPF	DL	384	m^3	$16m^3$/hr max capacity
GPF	AI	8	m^3	$16m^3$/hr max capacity

Readings in the wrong units (error flag -4)

We also discovered that all daily electricity active import readings were all extremely low for some participants. At the time of the initial investigation (August 2020) there were no participants with a maximum read between 85 and 2285 Wh; therefore we determined that any electricity active import (or export) meter with its highest reading less than 100 was erroneously reporting in kWh rather than in Wh as specified by the Smart Energy Code. This was verified by comparing the daily readings with the sum of the half-hours for the same day. Any readings that were deemed to be valid according to all other criteria were given the 'Elec in kWh' error flag and their data were multiplied by 1000 in the 'unit-corrected' column. Note that we set a limit of at least 30 readings in order to define a meter as recording in kWh.

Subsequent investigations in March 2021 revealed that some meters were correctly recording daily electricity readings in Wh up until the date of a meter replacement, at which the daily reads became approximately 1/1000th of the sum of the half-hourly reads. By 'approximately' we mean that after dividing the half-hourly sum by 1000 and rounding down, the result is within 1 of the daily read. For example, half-hourly sums between 5000 and 599 would be classed as approximately 1000 times bigger than a daily read between 4 and 6 (in order to handle rounding errors and slight read mismatch). In such cases, if a household has at least 5 rows with the daily and half-hourly sums in this situation (excluding daily reads of 0 which may be a different kind of error), all of such cases for the household are flagged as a unit error as above. Without checking all cases manually, a very small number of reads will be mis-flagged as unit errors or unit errors will be missed.

It is possible that some half-hourly sums are approximately 1/100th of the daily read, indicating readings in tens of Wh. This has been found to affect just a handful of meters at present, and it is left to researchers to decide how best to deal with daily and half-hourly sum mismatches in general. In most cases we believe that the sum of half-hours is more reliable than a daily read, but it depends on the individual meter, and not all days have complete valid half-hourly reads to sum. Note that if there are not the correct number of valid half-hourly reads taken at the right times then the half-hourly sum variable will be NA.

Zero Reads

Zero reads are not yet flagged as invalid although this may change in subsequent data releases. It has been found that some daily reads default to zero during British Summer Time (BST) which could be an obvious error to flag, but the validity of other zero reads is less clear. Considering mismatches between half-hourly sums and daily reads is advised when performing data quality analysis.

Flags for the daily and half-hourly sum match

The daily data table contains fields called “Elec_sum_match” and “Gas_sum_match” which give an error flag for how the sum of the half-hourly reads for that day compares with the daily read. They are described in the table below. It isn’t always possible to compare the sum and the daily readings; if any were reported in the wrong units, if any reads were taken at the wrong time, or if any of the reads were missing. By ‘any’ we mean the daily read and the 48 half-hourly reads (46 or 50 on clock change days).

Table 6: Electricity and gas ‘sum_match’ values and their meanings.

Code	*Meaning*
3	Daily electricity read originally recorded in kWh, match not tested due to rounding issues.
2	No meter: the (gas) meter does not exist in the DCC inventory
1	Daily read and half-hourly sum match: for electricity difference <= 1 Wh, for gas difference <= 1 L
0	Comparison not possible: do not have 48 valid half-hourly reads or daily read isn't valid
-1	Daily read and half-hourly sum are similar but don't match: for electricity 1 < difference <= 10 Wh, for gas 1 < difference <= 10 L
-2	Daily read and half-hourly read are neither similar nor match but are both valid

Data quality summary tables

There are two tables that give a summary of the data quality at 1) the individual read type level and 2) the participant level.

Read-type data quality summary

A read type is a combination of device type and schedule type, such as daily active electricity import or half-hourly reactive electricity export. There are 25 columns in the read-type data quality table. Each read type for each participant has its own row, which gives the number of readings with each error code, the start and end dates of the schedule (when we expect the earliest and latest readings to be), and some calculated columns for the percentage missing and valid etc. The details are given in the table below. Note that reads with error code 3 are not included in this table as they are an artifact of the data structure rather than genuine reads (see above for details).

Table 7: Read-type data quality summary: data fields. Error codes are explained above in the Error Flags section. Note that statistics for valid reads exclude valid reads recorded at the wrong time.

Field	*Description*	*Units*	*Class*	*Example value*
PUPRN	Pseudonymised participant identifier	NA	character	1VUXXXF1
deviceType	Device type: gas (GPF) or electricity (ESME) meter	NA	character	GPF
readType	Defined in Table 1	NA	character	AI
theoreticalStart	Earliest possible reading for the schedule	%Y-%m-%d	Date	2019-11-01
theoreticalEnd	Latest possible reading for the schedule	%Y-%m-%d	Date	2020-02-29
firstValidReadDate	Earliest date with a valid read (error flag 1 and Valid_read_time = TRUE)	%Y-%m-%d	Date	2018-11-26
lastValidReadDate	Latest date with a valid read (error flag 1 and Valid_read_time = TRUE)	%Y-%m-%d	Date	2020-05-31
daysRange	Schedule length = scheduleEnd - scheduleStart + 1	NA	numeric	100
maxPossReads	Maximum possible reads available (= daysRange for daily data, = 48 * daysRange for half-hourly)	NA	numeric	4800
percValid	Percentage of possible reads that are valid (error flag 1 and Valid_read_time = TRUE) rounded to 2 decimal places	NA	numeric	95.02

Field	*Description*	*Units*	*Class*	*Example value*
percValidOrUnitError	Percentage of possible reads that are valid or have a unit error (Valid_read_time = TRUE and error flag 1, or -4) rounded to 2 decimal places	NA	numeric	96.98
percMissing	Percentage of possible reads that are missing (error flag 0) rounded to 2 decimal places	NA	numeric	2.13
percError	Percentage of possible reads that are erroneous (error flag -1, -2, -3, -4 or Valid_read_time = FALSE) rounded to 2 decimal places	NA	numeric	3.04
valid	Number of valid readings taken at the right time (error flag 1 and Valid_read_time = TRUE)	NA	integer	96
validWrongTime	Number of valid readings taken at the incorrect time (error flag 1 and Valid_read_time = FALSE)	NA	integer	7
wrongUnits	Number of readings with electricity recorded in kWh (error flag -4)	NA	integer	3
missing	Number of missing readings (error flag 0)	NA	numeric	27
maxRead	Number of readings with the 'Max Read' error (flag -1)	NA	integer	4
highRead	Number of readings between the 'very high' and the 'max read' thresholds (error flag -2)	NA	integer	2
negative	Number of negative readings with (error flag -3)	NA	numeric	0
minValidRead	Minimum read of the valid reads (after unit-correction if necessary)	Wh (elec), m^3 (gas)	numeric	0
maxValidRead	Maximum read of the valid reads (after unit-correction if necessary)	Wh (elec), m^3 (gas)	numeric	302
meanValidRead	Mean of the valid reads (after unit-correction if necessary), 2 decimal places	Wh (elec), m^3 (gas)	numeric	43.21
medianValidRead	Median of the valid reads (after unit-correction if necessary)	Wh (elec), m^3 (gas)	numeric	46
sdValidRead	Standard deviation of the valid reads (after unit-correction if necessary), 2 decimal places	Wh (elec), m^3 (gas)	numeric	1.39

Participant-level summary

The second data quality summary table has one row per participant and includes additional information about the participant such as the region where they live and how many survey questions they answered. It also provides information about the start and end dates of each schedule and the number and percentage of reads that were valid for that schedule. There are 40 columns in this data table.

The read-related column names take the form "text_W_X_Y_Z" for electricity readings and "text_W_X_Z" for gas readings. Rather than explaining every single column, here is key to the variable component of the name:

- W = device type (either electricity meter ("ESME") or Gas Proxy Function i.e. gas meter ("GPF"))
- X = whether the read is half-hourly ("HH") or daily ("D")
- Y = active ("Act") or reactive ("React") power (electricity reads only)
- Z = import ("Im") or export ("Ex")

For example "NumValid_ESME_D_Act_IM" is the number of valid daily electricity active import readings. The text in the first part of the name is described in the table below.

Table 8: Participant data quality summary: data fields

Field (or field name format)	*Description*	*Units*	*Class*	*Example value*
PUPRN	Pseudonymised participant identifier	NA	character	1VUXXXF1
Region	Region in GB	NA	character	East Midlands
LSOA	Lower Super Output Area in GB	NA	character	E01015916
grid_cell	Grid cell for linking to climate data	NA	character	38_31
IMD_quintile	Index of Multiple Deprivation quintile (1 is most deprived, 5 is least deprived)	NA	integer	2
EPC_exists	TRUE if a record exists in the EPC dataset for the participant	NA	logical	TRUE
EPC_rating	EPC rating (original column name 'current_energy_rating') - a letter between A and G inclusive	NA	character	C

Field (or field name format)	*Description*	*Units*	*Class*	*Example value*
pilot_test_cell	Test cell (1 to 12) assigned to participants for the pilot study (testing recruitment approaches)	NA	integer	12
N_survey_ans	Number of questions answered in the survey (30 relevant to all, a further 9 may be relevant depending on other answers)	NA	integer	35
Perc_survey_ans			numeric	
invalidReadTimes_D	Number of rows in the daily data with the read at the wrong time (not at midnight)	NA	integer	1
Start_W_X_Y_Z	First valid read date for data type W_X_Y_Z	%Y-%m-%d	integer	2019-11-01
End_W_X_Y_Z	Last valid read date for data type W_X_Y_Z	%Y-%m-%d	Date	2020-02-29
NumValid_W_X_Y_Z	Number of valid reads (error code 1) for data type W_X_Y_Z	NA	Date	98
PercValid_W_X_Y_Z	Percentage of possible reads (using theoretical start and end dates rather than actual valid read start and end dates) that are valid (error code 1) for data type W_X_Y_Z	NA	integer	95.2