Smart meter consumption data: Technical documentation

Creation date	2020-08-25
Version	01
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_{version}.csv**: daily electricity and gas readings with some additional derived columns
- **SERL_smart_meter_hh_{version}.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_read_type_summary_{version}.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_{version}.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 *{version}* is the version of the data release, e.g. *v2020_08* for the data release created in August 2020.

This document is structured as follows: we start with some basic information about how the data was 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 (midnight - midnight).

Smart meter read types

Smart meter read types are defined by two variables: 'deviceType' and 'readType'. Together the 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

Smart meter data tables

Half-hourly data

The half-hourly data table has 20 columns. The fields are described in the table below. While the original data has 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 3.6.2 (2019-12-12) (code file *SERL_smart_meter_data_processing_v2020_08.R*).

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. Added or derived variables are shown in bold. 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*
PUPRN	Pseudonymised participant identifier	NA	character	1VUXXXF1
Read_date_effective	Date of read (same as date of Read_date_time unless read taken at midnight, then the previous day since data is for half hour on previous day)	%Y-%m-%d	Date	2019-11-01
Read_date_time	Time read taken	%Y-%m-%d %H:%M:%S	POSIXct, POSIXt	2019-11-02 00:00:00
HH	Half-hour identifier between 1 and 48 (NA if not on the half-hour)	NA	integer	48
Valid_read_time	FALSE if read time is not on the hour or half hour, otherwise TRUE	NA	logical	TRUE
Elec_import_exists	TRUE if electricity import meter exists in the inventory, otherwise FALSE	NA	logical	TRUE
Elec_act_imp_hh_Wh	Half-hourly electricity active import read	Wh	integer	109
Elec_act_imp_flag	Half-hourly electricity active import error flag	NA	numeric	-2
Elec_react_imp_hh_varh	Half-hourly electricity reactive import read	varh	integer	15
Elec_react_imp_flag	Half-hourly electricity reactive import error flag	NA	numeric	1
Elec_export_exists	TRUE if electricity export meter exists in the inventory, otherwise FALSE	NA	logical	FALSE
Elec_act_exp_hh_Wh	Half-hourly electricity active export read	Wh	integer	65
Elec_act_exp_flag	Half-hourly electricity active export error flag	NA	numeric	-4
Elec_react_exp_hh_varh	Half-hourly electricity reactive export read	varh	integer	14
Elec_react_exp_flag	Half-hourly electricity reactive export error flag	NA	numeric	2
Gas_exists	TRUE if gas meter exists in the inventory, otherwise FALSE	NA	logical	TRUE
Gas_hh_m3	Half-hourly gas import read	m^3	numeric	0.244
Gas_hh_Wh	Half-hourly gas import read in Wh using standard conversion, assuming calorific value = 39.5	Wh	numeric	2737.835
Gas_hh_kWh	Half-hourly gas import read in kWh using standard conversion, assuming calorific value = 39.5	kWh	numeric	2.737835

Field	Description	Units	Class	Example value
Gas_flag	Half-hourly gas import error flag	NA	numeric	0

Daily data

The daily data table has 22 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 3.6.2 (2019-12-12) (code file *SERL_smart_meter_data_processing_v2020_08.R*).

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 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 stating the number of valid half-hourly readings for each day (for electricity and gas), the sum of these half-hourly readings (if there were 48 valid reads), 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 5 in the Error Flags section. Note that reads are considered invalid if taken at the wrong time.

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
PUPRN	Pseudonymised participant identifier	NA	character	1VUXXXF1
Read_date_effective	Date of read	%Y-%m-%d	Date	2019-11-01
Read_date_time	Time and date of read	%Y-%m-%d	POSIXct, POSIXt	2019-11-01
Valid_read_time	TRUE if reading was at midnight, otherwise FALSE	NA	logical	TRUE
Valid_24h_read_flag	1 if this read and previous read both have Valid_read_time = TRUE, otherwise 0	NA	integer	TRUE

Field	*Description*	*Units*	*Class*	*Example value*
Elec_import_exists	TRUE if electricity import meter exists in the inventory, otherwise FALSE	NA	logical	TRUE
Elec_act_imp_d_Wh	Daily electricity active import read	Wh	integer	5839
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
Elec_act_imp_d_kWh	Unit_correct_elec_act_imp_d_Wh divided by 1000. Note that reads originally recorded in kWh will be integers, otherwise 3 decimal places.	kWh	numeric	5.839
Elec_act_imp_flag	Daily electricity active import error flag	NA	numeric	-2
N_elec_hh	Number of valid half-hourly elecricity active import readings available on this date excluding invalid read time data	NA	integer	48
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)	Wh	integer	5742
Elec_act_imp_sum_diff	Unit_correct_elec_act_imp_d_Wh - Elec_act_imp_hh_sum_Wh	Wh	numeric	0
Elec_sum_match	Error code for whether the sum of half-hourly electricity active import matches the daily electricity read	NA	numeric	1
Gas_exists	TRUE if gas meter exists in the inventory, otherwise FALSE	NA	logical	TRUE
Gas_d_m3	Daily gas read	m^3	numeric	8.214
Gas_d_kWh	Daily gas import read in kWh using standard conversion, assuming calorific value = 39.5	kWh	numeric	92.16628
Gas_flag	Daily gas import error flag	NA	numeric	2
N_gas_hh	Number of valid half-hourly gas import readings available on this date excluding invalid read time data	NA	integer	48
Gas_hh_sum_m3	Sum of half-hourly gas reads for this date (NA if there were not 48 valid reads)	m^3	numeric	8.763
Gas_sum_diff	Gas_d_m3 - Gas_hh_sum_m3	m^3	numeric	0.273
Gas_sum_match	Error code for whether the sum of half-hourly gas import matches the daily gas read	NA	numeric	0

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.

Table 4: Error flags and their meanings.

Flag	*Meaning*	*Details*
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
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 greater than 10^6 Wh (electricity) or 10^3 cubic metres (gas) but not a 'Max read'
-3	Negative	The read is negative (none found at this point)
-4	Elec in kWh	The electricity read was reported in kWh rather than Wh - details below
-5	Invalid read time	The read was taken at the incorrec time - overwrites all error flags except no meter exists

Very large readings (error flags -1 and -2)

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. There are also some very high readings, which, in order to be very cautious we define in this initial exploratory analysis stage as 10^6 Wh for electricity and 10^3 m^3 for gas.

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 investiation 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 it's highest reading less than 100 was erroneously reporting in kWh rather than in Wh as specified by the Smart Energy Code. 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 multipled 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.

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).

Table 5: 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 not 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.

Table 6: Read-type data quality summary: data fields. Error codes are explained above in the Error Flags section.

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	Schedule type: DailyLog, ActiveImport, Export, ReactiveImport, ReactiveExport, ReactiveExport. Note that all are half-hourly except DailyLog	NA	character	ActiveImport
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 code 1)	%Y-%m-%d	Date	2018-11-029
lastValidReadDate	Latest date with a valid read (error code 1)	%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 code 1) rounded to 2 decimal places	NA	numeric	95.02
percValidOrUnitError	Percentage of possible reads that are valid or have a unit error (error code 1 or -4) rounded to 2 decimal places	NA	numeric	96.98
percMissing	Percentage of possible reads that are missing (error code 0) rounded to 2 decimal places	NA	numeric	2.13
percError	Percentage of possible reads that are erroneous (error code -1, -2, -3, -4 or -5) rounded to 2 decimal places	NA	numeric	3.04
f_1	Number of readings with error code 1 (valid)	NA	integer	96
f_0	Number of readings with error code 0 (missing)	NA	numeric	27
f_minus1	Number of readings with error code -1 (Max read)	NA	integer	4
f_minus2	Number of readings with error code -2 (Very high but not max)	NA	integer	2

Field	*Description*	*Units*	*Class*	*Example value*
f_minus3	Number of readings with error code -3 (negative)	NA	numeric	0
f_minus4	Number of readings with error code -4 (Electricity recorded in kWh)	NA	integer	3
f_minus5	Number of readings with error code -5 (Incorrect read time)	NA	integer	7
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 39 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 7: 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
gridCell	Grid cell for linking to climate data	NA	character	38_31
imdQuintile	Index of Multiple Deprivation quintile (1 is most deprived, 5 is least deprived)	NA	integer	2
epcExists	TRUE if a record exists in the EPC dataset for the participant	NA	logical	TRUE
epcRating	EPC rating (original column name 'current_energy_rating') - a letter between A and G inclusive	NA	character	C
numSurveyAns	Number of questions answered in the survey (30 relevant to all, a further 9 may be relevant depending on other answers)	NA	integer	35
percSurveyAns	% of relevant questions answered in the survey (may be greater than 100% if question relevance was not possible to determine or if question skipping was not done correctly)	NA	numeric	98.1203
invalidReadTimes_D	Number of rows in the daily data with the read at the wrong time (not at midnight)	NA	integer	1
invalidReadTimes_HH	Number of rows in the half-hourly data with the read at the wrong time (not on the hour or half hour)	NA	integer	3
Start_W_X_Y_Z	First valid read date for data type W_X_Y_Z	%Y-%m-%d	Date	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

Field (or field name format)	*Description*	*Units*	*Class*	*Example value*
NumValid_W_X_Y_Z	Number of valid reads (error code 1) for data type W_X_Y_Z	NA	integer	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	numeric	95.2