# Dataset of smart heat and water meter data with accompanying building characteristics

Schaffer, Markus; Veit, Martin; Marszal-Pomianowska, Anna; Frandsen, Martin; Pomianowski, Michal Zbigniew; Dichmann, Emil; Sørensen, Christian Grau; Kragh, Jesper

*Publication date:*
2023

*Document Version*
Other version

# Dataset of smart heat and water meter data with accompanying building characteristics

Markus Schaffer[1], Martin Veit[1], Anna Marszal-Pomianowska[1]*, Martin Frandsen[1], Michal Zbigniew Pomianowski[1], Emil Dichmann[2], Christian Grau Sørensen[2], Jesper Kragh[2]
[1]Department of the Built Environment, Aalborg University, Aalborg, Denmark
[2]Department of the Built Environment, Aalborg University, Copenhagen, Denmark
*ajm@build.aau.dk

## Abstract

The data presented were collected from 34884 commercial smart heat meters and 10765 commercial smart water meters with a span of up to 5 years (2018 - 2022), all located in buildings within the Aalborg Municipality. In addition, building characteristics from the Danish Building and Dwelling Register (BBR) and Energy Performance Certificate (EPC) input data were collected for each building, resulting in up to 86 different characteristics per building. All smart meter data was processed using an established method to obtain equidistant data without erroneous values. The building characteristics derived from the EPCs were filtered based on rule sets to increase the quality of the data. All data can be linked. It is expected that researchers in the built environment, district heating and water sectors will find such a dataset of great value.

| | |
|---|---|
| **Subject** | Civil and Structural Engineering |
| **Specific subject area** | Hourly smart heat and water meter data from buildings. Building characteristics. |
| **Type of data** | Table<br>Database |
| **How the data were acquired** | The hourly smart heat meter data was acquired via commercial smart heat meters (Kamstrup MULTICAL 402, Kamstrup MULTICAL 403, Kamstrup MULTICAL 603)<br>The hourly smart water meter data was acquired via commercial smart water meters (MULTICAL 21/ flowIQ 210x)<br>The statistical building characteristics were collected from the Danish Building and Dwelling Register (BBR)<br>The detailed building characteristics were collected from the input data of Danish Energy Performance Certificates (EPCs) |

| | |
|---|---|
| **Data format** | Raw<br>Filtered |
| **Description of data collection** | The local district heating and water company collected the hourly smart heat and water meter data for billing purposes. The statistical building characteristics were collected via a publicly available API. The detailed building characteristics were collected from a database at Aalborg University. All building characteristics were collected on the condition that energy and/or water consumption data were available for the building. |
| **Data source location** | Aalborg Municipality<br>Denmark<br>57.053, 9.924 |
| **Data accessibility** | Part of the data was originally collected for billing purposes (hourly data from smart heat and water meters) and made available to the authors for scientific purposes via a data use agreement on the legal basis of GDPR article 89. The data were anonymised by the researchers. However, as the data can potentially be deanonymised in combination with the building characteristics through a backward search in the public Danish Building and Dwelling Register, the data is considered personal data subject to the GDPR.<br>Researchers interested in using the data should contact the corresponding author and are then required to complete a joint declaration that the data sharing is lawful. It should be noted that for researchers outside the European Union, possible additional requirements apply in accordance with applicable Danish and European law. Once the agreement has been approved, the data can be accessed via an API, which requires authentication via edugain. |
| **Related research article** | M. Schaffer, J.E. Vera-Valdés, A. Marszal-Pomianowska, Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses, (2023). [1] |

# 1 Value of the data

- This dataset provides an unprecedented amount of data, particularly in conjunction with accompanying building features at a high level of detail. The easy and curated accessibility of the data makes it usable for both small and large-scale investigations.
- The data can be of great value to research in the built environment and the district heating and water sectors. It provides countless opportunities for data-driven research and validation of models.
- The data can be used to develop new data-driven methods to gain insight into the energy use of buildings. It can also be used to detect faults in buildings. In addition, it is valuable for validating urban building energy models or automatically deriving building characteristics based on energy use data.

# 2 Objective

The dataset was originally created to support research within the FOREFRONT project. It has now been decided to make this data generally available, recognising its high value to other researchers (with the aforementioned restrictions). Making it more widely available will increase the reproducibility of the research carried out by Schaffer et al., 2023 [1] and any future research carried out using this data.

# 3 Data description

The data is structured within six tables in a database. An entity relationship diagram is shown in Figure 1. All data can be related, which is the core idea of the whole database. For all processed data, the meter ID is unique and can be used as an identifier. For the raw smart meter data, it should be noted that there may be meters that are incorrectly assigned to two customers, so the uniqueness of the ID is not guaranteed for the raw data. The customer ID can be used to link Smart Heat Meter (SHM) and Smart Water Meter (SWM) data. It should be noted that a customer can have one or more meters. For this reason, there may be duplicate entries in the Danish Building and Dwelling Register (BBR) data, differing only in the meter ID, e.g., if a customer has one SHM but two SWM, then there are two entries, identical except for the SWM ID (both entries have the same SHM ID). For the BBR data, due to the dependency on the data period, there may be several identical entries for the same building, e.g., one for the SWM data, one for the SHM data, or several for the SHM data if the SHM data has several periods. Figure 2 gives an overview of the number of meters for which the respective data (processed data from SHM and SWH data) are available in the database. In the following, each table is described separately.

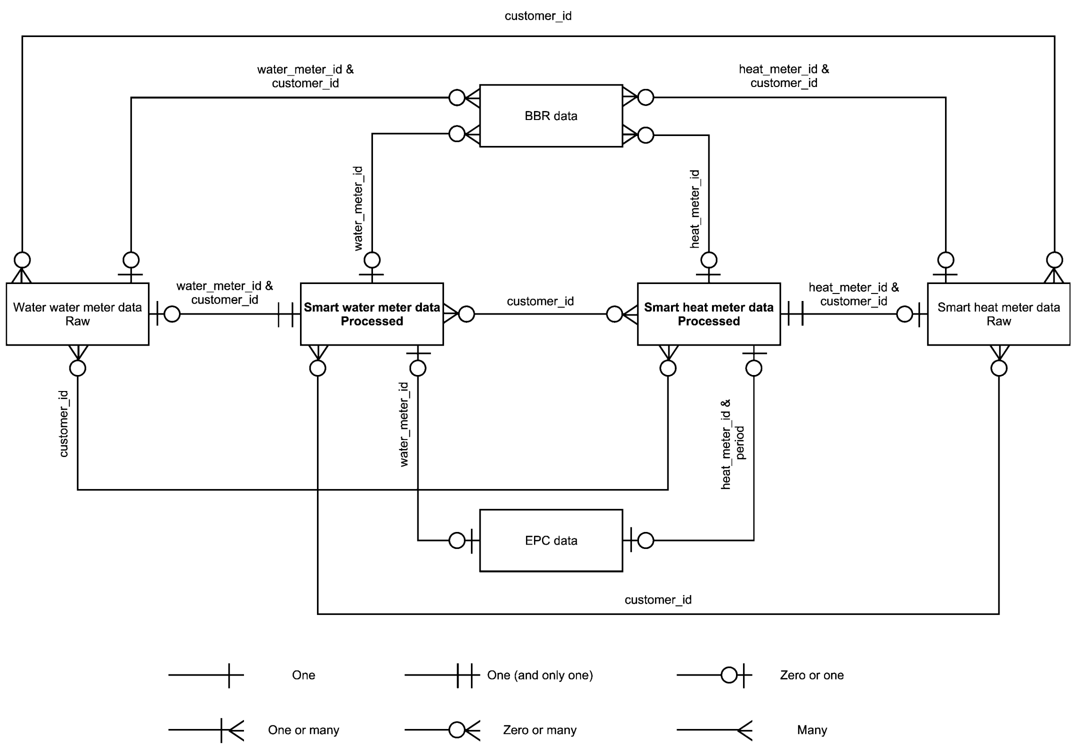


**Figure 1 Entity relationship diagram for the database**

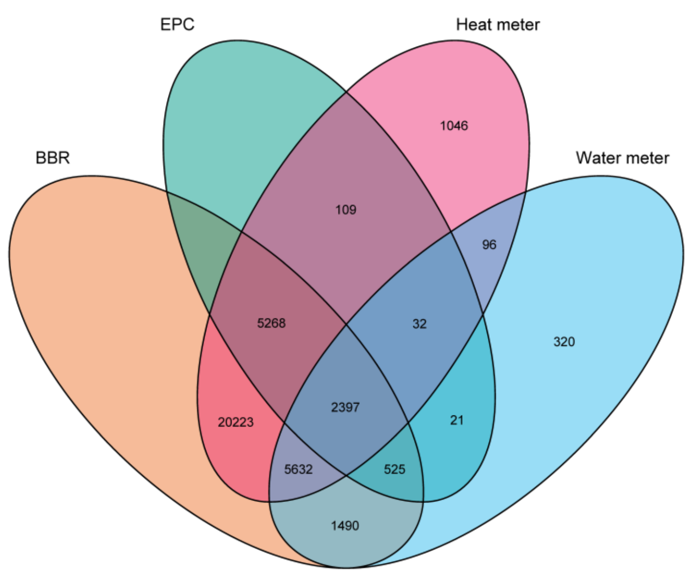


**Figure 2 Meter ID and customer ID based number of meters available in the respective group based on the processed data.**

## 3.1 Hourly smart heat meter data

### 3.1.1 Raw data

This table contains the data as collected by the SHMs installed in the respective buildings in the municipality of Aalborg, Denmark. An overview of all columns included in this dataset is given in Table 1. The data span from the beginning of 2018 to the end of 2022 (with different lengths for each building) and contain data from a total of 34884 SHMs (9.46e+08 rows). Data from a building may not be complete, i.e. a building may have data from 2018 and 2020 but no data from 2019. The data has not been processed in any way other than the removal of redundant columns of units of measurement to reduce the amount of storage space required. As the data is not processed, it is not equidistant hourly as the SHMs have a temporal accuracy of ±30 minutes around the full hour. In addition, the original data were delivered to the researchers with a timestamp in local time (CET/CEST) but without any time zone information. Consequently, the time may be incorrect at the end of summertime, as the two 'overlapping' hours at 03:00 could not be distinguished. The data contain missing values due to errors in the transmission infrastructure used to collect the data.

### 3.1.2 Processed data

The processed data table contains the processed data from the SHMs. It contains data from 34795 SHMs (9.33e+08 rows), and an overview of all available columns is given in Table 1. These data are equidistant, have no erroneous values, and missing ones have been imputed. The processing used is described in detail in Section 4.1. Figure 3 shows the number of SHMs available for the different years of the data period.

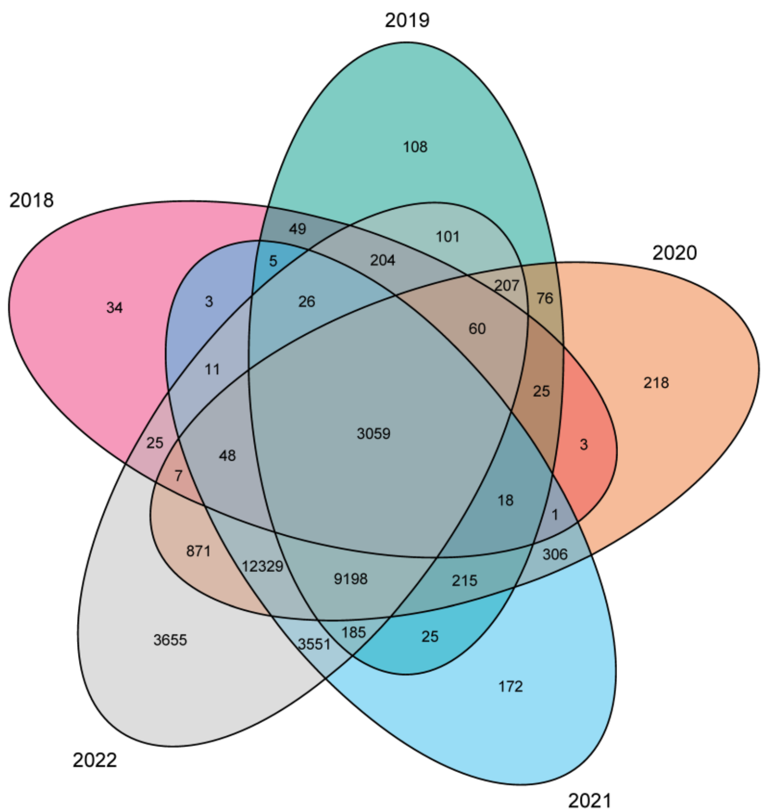


**Figure 3 Number of processed SHMs available across the different years of the data period.**

## 3.2 Smart water meter data

### 3.2.1 Raw data

This table contains the data as collected via the SWMs installed in the respective buildings in Aalborg Municipality, Denmark. An overview of all columns included in this dataset is given in Table 2. The data covers the period from the beginning of May 2021 to the end of 2022 (with different lengths for each building) and contains in total data from 10765 SWMs (7.19e+07 rows). The data has not been processed in any way other than removing redundant columns containing units of measurement to reduce the amount of storage required. As the data is not processed, it is not equidistant hourly as the SWMs have a time accuracy of ±30 minutes around the full hour. The data have been supplied with UTC timestamps, so unlike the SHM data, the timestamp is always correct.

### 3.2.2 Processed data

The processed data table contains the processed data from the SWMs. It contains data from 10510 SWMs (7.04e+07 rows), and an overview of all available columns is given in Table 2. These data are equidistant, have no erroneous values, and missing ones have been imputed. The processing used is described in detail in Section 4.2.

## 3.3 Statistical building characteristics (BBR)

For each building for which either SHM or SWM data are available in this dataset, the corresponding data from the BBR have been collected where possible. This publicly available database in Denmark contains information on every building in Denmark and is operated by the Danish Customs and Tax Administration. An overview of the available columns is given in Table 3.

## 3.4 Detailed building characteristics (EPC)

For each building for which either SHM or SWM data are available in this dataset, the input data from the corresponding energy performance certificate (EPC), if available, were collected and processed from the EPC database developed by Brøgger and Wittchen, 2016 [2] and hosted at Aalborg University. An overview of the available columns is given in Table 4. The processing used to derive the data is described in Section 4.4.

# 4 Experimental design, materials and methods

## 4.1 Smart heat meter data processing

The SHM data were obtained by the authors from the local utility company as .csv files. As mentioned above, the readings were provided in local time (CET/CEST) without any time zone information. As the dataset is similar to the one described in detail by Schaffer et al., 2022 [3], a similar cleaning and imputation framework was applied to obtain equidistant data without erroneous or missing values. The only difference to the framework described in Schaffer et al., 2022 [3] is that, due to the long data period and the higher uncertainty in data quality, it was tested that there were at least 8584 hours of data per year and per meter (approximately 2% of missing data). If this threshold was exceeded, only the year in question was excluded. Thus, an SHM may have data in non-consecutive years in the processed data. Consequently, these data sequences can be considered as separate data. For this reason, the period column (Table 1) has been introduced. This column, starting with one, indicates whether the data of the SHM are from a different sequence, i.e. if an SHM has data in 2018 and 2020-2022 but no data in 2019, the period column is 1 for all data in 2018 and 2 for all data in 2020-2022.

In addition to this basic data treatment, the SPMS method developed by Schaffer et al., 2023 [4] was applied to energy use. SPMS was developed to reduce the error introduced by rounding the raw cumulative energy data to integer values. The result of this process is available as a separate column (heat_energy_kwh_spms) in the processed data (Table 1).

## 4.2 Smart water meter data processing

The authors obtained the SWM data from the local utility company as .csv files. The data were provided with readings in UTC. Given the same nature of the data (cumulative and approximately hourly), the same cleaning and imputation framework as for the SHM data was used to process the SWM data. However, given the shorter data period compared to the SHM, the threshold for missing values was set at 2% for each SWM dataset individually to account for the different lengths of the datasets.

## 4.3 BBR data processing

The address was the only customer information provided by the utility company to link SHM and SWH data to a building/unit. It was unclear whether the address referred to a unit (e.g. an apartment) or a building (e.g. an apartment building). This information is used to retrieve the building characteristics from the BBR database. In order to prevent incorrect information from influencing the retrieval of building characteristics, the address information provided was treated with the Address Cleaning API, which is part of the Danish Address Web API (DAWA) [5]. This API can translate unstructured addresses with possible misspellings into official addresses. In addition to the address information, the API returns the certainty of the match expressed in three levels, A - identical match, B - certain match and C - uncertain match. Only results with a confidence of A or B are considered valid. As the address cleaning API distinguishes between unit and building addresses, all addresses were initially treated as unit addresses, and only addresses with a certainty of C were subsequently treated as building addresses. Addresses for which neither a unit nor a building address could be found with high confidence (level A or B) were excluded.
The BBR information was obtained through Denmark's Address Web API (DAWA) [5]. Information about a unit and its building could be obtained directly through the API. For the SHMs where only a building address was available, the 'access address id' had to be retrieved via the address before information about the building could be obtained. In both cases, more than one BBR record may be obtained, for example, if two or more units/buildings have the same address. In order to allow for a data structure where an SHM can be linked to zero or one BBR record, cases where more than one record was obtained were considered invalid and consequently not included in the database. All nominal values were translated to human understandable terms in English.

As the main objective was to establish essential building characteristics for as many SHMs as possible, only mandatory BBR information was considered for the dataset. Such mandatory information must be provided by the building owner and is, therefore, subject to uncertainty. Recently, however, the quality of the data was investigated [6] and it was concluded that the overall data quality is high and that the data quality has improved from 2000 to 2013.

## 4.4 EPC data processing

To link the available EPC data from the EPC database developed by Brøgger and Wittchen, 2016 [2] and hosted at Aalborg University with the SHM and SWM data, the same 'cleaned' addresses were used as for the BBR data (Section 4.3). Given the sheer amount of information available in the EPCs, it was decided to focus mainly on data from five aspects:

- Building envelope
- Domestic hot water (DHW)

- Ventilation
- Heating
- Internal heat gains

The data quality of the Danish EPC has been heavily criticised in the past, as random checks have revealed errors in 20-30% of all EPCs [7]. For this reason, the cleaning framework developed by Brøgger, 2019 [7] was applied. However, this framework was originally developed for the purpose of energy modelling of the building stock. Therefore, some criteria have been adapted, and some have been added to better fit the purpose of this dataset. All quality assurance criteria used are listed in Appendix AAppendix AAppendix A.

After the cleaning step, the information obtained was aggregated to obtain the same building characteristics for each building where information was available. The resulting columns, including a description of how they were calculated, are shown in Table 4. Only results where an EPC record could be clearly linked to one building were considered. Furthermore, only valid EPCs were considered. Validity was defined as the EPC being valid (no more than 10 years old) at least on the first day of the data period. For SHM data, each period was considered separately. Thus, if a SHM has two periods, one period may have EPC information available, and the other may not, or the information may differ between the periods. In addition, several EPCs can be valid simultaneously, as EPCs are not invalidated when a new EPC is issued. For example, if a house is sold, an EPC is issued, the house is renovated, and a new EPC is issued. If two EPCs are valid for an SHM or SWM, the information from the most recent EPC was used. Furthermore, if an EPC was issued during the data period of the respective SHM or SWM, all EPCs are considered invalid, as it is assumed that the building has been renovated and, therefore, the data represent two different building conditions.

# 5 CRediT author statement

**Markus Schaffer:** Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing **Martin Veit:** Methodology, Software, Data Curation, Writing - Review & Editing **Anna Marszal-Pomianowska:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition **Martin Frandsen:** Methodology, Writing - Review & Editing **Michal Zbigniew Pomianowski:** Methodology, Writing - Review & Editing **Emil Dichmann:** Software, Data, Curation Writing - Review & Editing **Christian Grau Sørensen:** Software, Data Curation, Writing - Review & Editing **Jesper Kragh:** Data Curation, Writing - Review & Editing

# 6 Acknowledgements

This work was funded by the Independent Research Fund Denmark under FOREFRONT project (0217-00340B). The authors would like to thank Aalborg Forsyning, notably Kenneth Faarkrog Kristensen, who kindly provided the original data.

# 7 Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

# 8 References

[1] M. Schaffer, J.E. Vera-Valdés, A. Marszal-Pomianowska, Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses, (2023).

[2] M. Brøgger, K.B. Wittchen, Energy Performance Certificate Classifications Across Shifting Frameworks, Procedia Eng. 161 (2016) 845–849. https://doi.org/10.1016/J.PROENG.2016.08.727.

[3] M. Schaffer, T. Tvedebrink, A. Marszal-Pomianowska, Three years of hourly data from 3021 smart heat meters installed in Danish residential buildings, Sci Data. 9 (2022) 420. https://doi.org/10.1038/s41597-022-01502-3.

[4] M. Schaffer, D. Leiria, J.E. Vera-Valdés, A. Marszal-Pomianowska, Increasing the accuracy of low-resolution commercial smart heat meter data and analysing its error, in: 2023 European Conference on Computing in Construction 40th International CIB W78 Conference, 2023.

[5] The Danish Agency for Data Supply, Denmark's Address Web API, (2022). https://dawadocs.dataforsyningen.dk/.

[6] SAS Institute, Datakvaliteten af ejerboliger i BBR, 2015. https://ki.bbr.dk/file/571501/datakvalitet-af-ejerboliger-i-bbr.pdf (accessed February 17, 2023).

[7] M. Brøgger, Building stock energy modelling, PhD, Aalborg University, 2019. https://doi.org/10.5278/vbn.phd.eng.00077.

[8] Trafik- Bygge- og Boligstyrelsen, Bygningsreglementets vejledning om korrektioner til 10 pct.-reglen for dagslys, Bygningsreglementets Vejledning Om Korrektioner Til 10 Pct.-Reglen for Dagslys. (2019) 25.

**Table 1 Description of the raw and processed smart heat meter (SHM) data**

| Column name | Raw data | Processed data | Description | Unit |
|---|---|---|---|---|
| customer_id | ✓ | ✓ | Hashed customer ID. Unique for every customer of the utility company. | - |
| heat_meter_id | ✓ | ✓ | Hashed meter ID. Unique for every SHM (guaranteed unique only in the processed data) | - |
| period | | ✓ | As a meter can have data for non-consecutive years period indicates if the data of one meter is continuous or from two or more separated years. A period is thereby an integer ranging from 1 to n. | |
| reading_time | ✓ | | Original reading time of the SHM given in local time (CET). To be noted, the time is saved in the database correctly parsed with the time zone. However, originally the time was supplied without a time zone. Thus, the time can be incorrect when the daylight-saving time ends. | - |
| time_rounded | ✓ | ✓ | Equidistant timesteps as a result of the data processing given in local time (CET). | - |
| heat_energy_kwh | ✓ | ✓ | Cumulative heat energy deposited. Raw values are rounded down to inter-values. | kWh |
| heat_energy_kwh_demand | | ✓ | Calculated hourly energy use | kWh |
| heat_energy_kwh_spms | | ✓ | By SPMS [4] treated energy use data. This reduces the rounding error introduced by the rounding down of the original data. | kWh |
| volume_m3 | ✓ | ✓ | The cumulative volume of district heating water passed through the SHM - measured at the supply Raw values are rounded down to inter values. | m³ |
| volume_m3_demand | | ✓ | Calculated hourly volume use | m³ |
| flow_x_temp_supply_m3C | ✓ | ✓ | The cumulative volume flow of the supply multiplied by the supply temperature. Raw values are rounded down to inter-values. | m³°C |
| flow_x_temp_supply_m3C_demand | | ✓ | The demand value of the volume flow of the supply multiplied by the supply temperature. | |
| flow_x_temp_return_m3C | ✓ | ✓ | The cumulative volume flow of the supply multiplied by the return temperature. Raw values are rounded down to inter-values. | m³°C |
| flow_x_temp_return_m3C_demand | | ✓ | The demand value of the volume flow of the supply multiplied by the return temperature. | |
| was_missing | | ✓ | Binary column indicating if a value was imputed | |

| Column name | Raw data | Processed data | Description | Unit |
|---|---|---|---|---|
| supply_temp_C | ✓ | | Instantaneous supply temperature at the time of reading (reading_time) | °C |
| return_temp_C | ✓ | | Instantaneous return temperature at the time of reading (reading_time) | °C |
| supply_flow_m3 | ✓ | | Instantaneous supply flow at the time of reading (reading_time) | $m^3/h$ |
| time_counter_h | ✓ | | Number of hours the SHM has been in operation | h |
| heat_effect_kw | ✓ | | Current deposited heating effect at the time of reading (reading_time) – not recorded for all meter types | kW |
| meter_type | ✓ | | SHM type | - |

**Table 2 Description of the raw and processed smart water meter (SWM) data**

| Column name | Raw data | Processed data | Description | Resolution |
|---|---|---|---|---|
| customer_id | ✓ | ✓ | Hashed customer ID. Unique for every customer of the utility company. One customer can have multiple SWMs | - |
| water_meter_id | ✓ | ✓ | Hashed meter ID. Unique for every SHM (guaranteed unique only in the processed data) | - |
| reading_time | ✓ | | Original reading time of the SWM given in local time (CET). In contrast to the smart heat meter data, this data was originally saved in UTC, and thus no issue with the daylight-saving time exists. | - |
| time_rounded | ✓ | ✓ | Equidistant timesteps as a result of the data processing given in local time (CET). | - |
| water_volume_m3 | ✓ | ✓ | Cumulative water volume | m3 |
| water_volume_demand_m3h | | ✓ | Hourly usage of water | m3 |
| was_missing | | ✓ | Binary column indicating if a value was imputed | |

**Table 3 Description of data derived from the Danish Building and Dwelling Register**

| Column name | Type | Description |
|---|---|---|
| heat_meter_id | - | Unique meter ID, which functions as the key to link the data to the smart heat meter data |
| Water_meter_id | | |
| unit_type_code | nominal | Type of unit, such as a single-family house, apartment etc. |
| unit_housing_type_code | nominal | Information if the unit is a residential apartment, mixed-used, a single room etc. |
| unit_total_area | float | Total area of the unit |
| unit_residential_area | float | Total residential area of the unit |
| unit_business_area | float | Total business area of the unit |
| unit_nr_room | integer | Number of rooms in the unit |
| unit_toilet_pos_code | nominal | Information if the toilet is positioned inside the unit or outside the unit. |
| unit_bath_pos_code | nominal | Information if a bathroom exists and if the bathroom is positioned inside the unit or outside the unit. |
| unit_kitchen_pos_code | nominal | Information if a kitchen exists and if the kitchen is positioned inside the unit or outside the unit. |
| unit_energy_code | nominal | Information about which voltage of electricity is available in the unit and if gas is available. |
| unit_nr_business_room | integer | Number of business rooms per unit |
| unit_other_area | float | Total area which is neither business nor residential |
| unit_rent_status_code | nominal | Information if the unit is used by the owner or rented out. |
| unit_heating_code | nominal | Type of heating available in the unit |
| unit_heating_agent_code | nominal | Type of heating agent used for heating the unit |
| unit_sup_heating_code | nominal | Type of supplementary heating available in the unit |
| unit_nr_toilet | integer | Number of toilets in the unit |
| unit_nr_bathroom | integer | Number of bathrooms in the unit |
| bldg_type_code | nominal | Same as unit_type_code but on the building level |
| bldg_nr_units_w_kitchen | integer | Number of units with kitchen in the building |
| bldg_nr_units_wo_kitchen | integer | Number of units without a kitchen in the building |
| bldg_constrcution_year | integer | Construction year of building |
| bldg_conversion_year | integer | Year of renovation of the building |
| bldg_ext_wall_mat_code | nominal | Type of external façade cladding |
| bldg_roof_mat_code | nominal | Type of roof cladding material |

| Column name | Type | Description |
|---|---|---|
| bldg_sup_ext_wall_mat_code | nominal | Type of supplementary external façade cladding |
| bldg_sup_roof_mat_code | nominal | Type of supplementary roof cladding material |
| bldg_total_area | float | Building total area |
| bldg_residential_area | float | Building residential area |
| bldg_business_area | float | Building business area |
| bldg_developed_area | float | Developed area of the building |
| bldg_nr_floor | integer | Number of floors in the building |
| bldg_floor_code | nominal | Information about the floors, e.g., if the building has double high storeys or deviating floors. |
| bldg_heating_code | nominal | Same as unit_heating_code but on the building level |
| bldg_heating_agent_code | nominal | Same as unit_heating_agent_code but on the building level |
| bldg_sup_heating_code | nominal | Same as unit_sup_heating_code but on the building level |
| bbr_resolution | nominal | Information on whether the address could be attributed to a unit or a building. If the address could only be linked to a building, information about the unit are missing. |

**Table 4 Description of data derived from the Danish Energy Performance Certificate**

| | Column name | Unit/Type | Description |
|---|---|---|---|
| General information | heat_meter_id | nominal | Unique meter ID, which functions as the key to link the data to the smart heat meter data |
| | water_meter_id | | |
| | period | nominal | As a meter can have data for non-consecutive years period indicates if the data of one meter is continuous or from two or more separated years. A period is thereby an integer ranging from 1 to n. |
| | valid_from | | |
| | valid_to | | |
| General building characteristics | bbr_use_code | nominal | Use code as defined in the Danish Building and Dwelling Register (translated) |
| | total_heated_floor_area | m² | The total heated floor area of the building |
| | heated_commercial_area | m² | Commercial area of the building |
| | height | m | |
| | floor_count | - | Number of floors of the building |
| | heat_capacity | W/(m²K) | Simplified heat capacity of the building according to DS/INF 418-2:2014 (or an earlier version if the data is based on an EPC from before 2014) per unit gross area |
| Opaque envelope | opaque_heatloss_kelvin | W/K | Total heat losses through the opaque envelope per Kelvin calculated as follows: $$\sum_{n=1}^{i} area_n \times u - value_n \times temperature\ factor_n \quad (1)$$ |
| Opaque envelope | opaque_heatloss_total | W | Total heat losses through the opaque envelope, taking the dimensioning temperature into account, calculated as follows: $$\sum_{n=1}^{i} area_n \times u\ value_n \times temperature\ factor_n \times (\text{dim. int. temp.} - \text{dim. ext. temp.}) \quad (2)$$ The dimensioning temperatures are thereby calculated based on the Danish standard DS 418:2011. Standard values are thereby 20°C for the interior, 30°C interior temperature for a floor with floor heating, -12°C for the exterior, and 10°C for exterior elements against soil deeper than 2m. |

| | Column name | Unit/Type | Description |
|---|---|---|---|
| Window north | window_heatloss_north_kelvin | W/K | Heat losses per Kelvin through all windows facing north (orientation > 315° OR orientation <= 45°), calculated as: $$\sum_{n=1}^{i} nr\ of\ windows_n \times area_n \times u - value_n \times temperature\ factor_n \quad (3)$$ |
| | window_heatloss_north_total | W | Total heat losses through all windows facing north (orientation > 315° OR orientation <= 45°), taking the dimensioning temperature into account is calculated as: $$\sum_{n=1}^{i} nr\ of\ windows_n \times area_n \times u\ value_n \times temperature\ factor_n \times (\text{dim. int. temp.} - \text{dim. ext. temp.}) \quad (4)$$ The dimensioning temperatures are thereby calculated based on the Danish standard DS 418:2011. Standard values are thereby 20°C for interior, 30°C interior temperature for floor with floor heating, -12°C for exterior, and 10°C for exterior for elements against soil deeper than 2m. |
| Window north | window_solar_north | - | Total solar factor of all windows facing north (orientation > 315° OR orientation <= 45°), calculated as: $$\sum_{n=1}^{i} nr\ of\ windows_n \times area_n \times g - value_n \times glass\ share_n \times \text{shading factor} \quad (5)$$ Whereby the shading factor was calculated from the angles to shading objects of each window based on the simplified method stated in [8]. For objects shading from the side as well as overhang, an infinite height respectively length was assumed. It is to be noted that the shading from the wall thickness could not be considered as the simplified method is based on the wall thickness, which is not an input for EPCs. |

| | | | |
|---|---|---|---|
| Window east | window_heatloss_east_kelvin | W/K | Heat losses per Kelvin through all windows facing east (orientation > 45° AND orientation <= 135°), calculated as stated in Equation 3. |
| | window_heatloss_east_total | W | Total heat loss through all windows facing east (orientation > 45° AND orientation <= 135°), taking the dimensioning temperature into account, is calculated as stated in Equation 4. |
| | window_solar_east | - | Total solar factor of all windows facing east (orientation > 45° AND orientation <= 135°), calculated as stated in Equation 5 |
| Window south | window_heatloss_south_kelvin | W/K | Heat losses per Kelvin through all windows facing south (orientation > 135° AND orientation <= 225°), calculated as stated in Equation 3. |
| | window_heatloss_south_total | W | Total heat loss through all windows facing east (orientation > 135° AND orientation <= 225°), taking the dimensioning temperature into account, is calculated as stated in Equation 4. |
| | window_solar_south | - | Total solar factor of all windows facing south (orientation > 135° AND orientation <= 225°), calculated as stated in Equation 5 |
| Window west | window_heatloss_west_kelvin | W/K | Heat losses per Kelvin through all windows facing west (orientation > 225° AND orientation <= 315°), calculated as stated in Equation 3. |
| | window_heatloss_west_total | W | Total heat loss through all windows facing east (orientation > 225° AND orientation <= 315°), taking the dimensioning temperature into account, is calculated as stated in Equation 4. |
| | window_solar_west | - | Total solar factor of all windows facing west (orientation > 225° AND orientation <= 315°), calculated as stated in Equation 5 |
| Skylight | skylight_heatloss_kelvin | W/K | Heat losses per Kelvin through all skylights were calculated as stated in Equation 3. |
| | skylight_heatloss_total | W | Total heat loss through all skylights taking the dimensioning temperature into account, is calculated as stated in Equation 4. |
| | skylight_solar | - | The total solar factor of all skylights was calculated as stated in Equation 5 |
| Thermal bridge | thermal_bridge_kelvin | W/K | Total heat losses through thermal bridges, calculated as follows: $$\sum_{n=1}^{i} length_n \times u - value_n \times temperature\ factor_n \qquad (6)$$ |
| Thermal bridge | thermal_bridge_total | W | Total heat losses through thermal bridges, taking the dimensioning temperature into account, calculated as follows: $$\sum_{n=1}^{i} length_n \times u\ value_n \times temperature\ factor_n \times (\text{dim. int. temp.} - \text{dim. ext. temp.}) \qquad (7)$$ |

| | | | | |
|---|---|---|---|---|
| Domestic hot water tank | dhw_tank_volume | l | Total domestic hot water tank volume calculated as: $$\sum_{n=1}^{i} nr\ of\ tanks_n \times volume_n$$ It is, however, to be noted that the Danish EPC calculation method is insensitive to the tank volume. For this reason, many buildings have a total share of domestic hot water covered by the domestic hot water tank larger than zero with a 0l tank volume. | **8** |
| | dhw_tank_heat_loss | W/K | Total heat losses from domestic hot water tanks, calculated as follows: $$\sum_{n=1}^{i} nr\ of\ tanks_n \times heat\ loss_n \times temperature\ factor_n$$ | **9** |
| Domestic hot water tank | dhw_tank_sup_temp | °C | Required supply flow temperature from the central heating system to the domestic hot water tank calculated as follows: $$\frac{\sum_{n=1}^{i} supply\ temperature_n}{n}$$ Due to the above-mentioned fact that a large share of EPCs have a tank volume of 0l, the tank volume is not considered for averaging. | **10** |
| | dhw_tank_share | - | The total share of domestic hot water covered by the domestic hot water tanks. Calculated as: $$\sum_{n=1}^{i} share\ of\ consumption_n$$ | **11** |
| | dhw_tank_el_support_code | nominal | A factor indicating whether the domestic hot water tank has electrical heating. The factor has three levels:<br>• *None*: no electric heating<br>• *Always*: electric heating is always available<br>• *Summer:* electric heating is only available in summer available<br>• No tank<br>The value was derived based on the maximum number of tanks with the respective electric heating possibility. (The volume could be due to the above problem, that many EPC have erroneously a 0l tank, not be used) | |

| | | | |
|---|---|---|---|
| Domestic hot water | dhw_average_consumption | l/year | Domestic hot water demand calculated as: $$heated\ dwelling\ area\ \times \text{average DHW consumption} \quad (12)$$ |
| | dhw_temperature | °C | Domestic hot water temperature |
| | dhw_pipes | W | Total heat losses through DHW pipes calculated as follows: $$\sum_{n=1}^{i} length_n \times u - value_n \times temperature\ factor_n \quad (13)$$ |
| Internal gains | gains_people | W | Total heat gains from occupants, calculated as follows: $$\sum_{n=1}^{i} area_n\ \times occ\ heat\ gains\ \text{per area}_\text{n} \quad (14)$$ |
| | gains_device | W | Total heat gains from appliances inside usage hours, calculated as follows: $$\sum_{n=1}^{i} area_n\ \times appliances\ heat\ gains\ per\ area_n \quad (15)$$ |
| | gains_device_outside | W | Total heat gains from appliances outside usage hours calculated as stated in Equation 15 |
| Heating system | heating_supply_temp | °C | Supply temperature of the heat distribution system |
| | heating_return_temp | °C | Return temperature of the heat distribution system |
| | heating_pipes | W | Total heat losses through heating pipes, calculated as: $$\sum_{n=1}^{i} length_n \times u - value_n \times temperature\ factor_n \quad (16)$$ |
| | heating_type_code | nominal | Plant type:<br>• 1: Single-circuit system<br>• 2: Double circuit system (or parts of the installation are single circuit, and these are equipped with local mixing devices) |

| | | | | |
|---|---|---|---|---|
| Ventilation winter | vent_nat_winter | l/s | Total natural ventilation in winter, calculated as follows:: $$\sum_{n=1}^{i} area_n \times ventilationflowperarea_n \times \text{usagefactor}_n$$ | **17** |
| | vent_mech_winter | l/s | Total mechanical ventilation in winter, calculated as follows: $$\sum_{n=1}^{i} area_n \times ventilation\ flow\ per\ area_n \times usage\ factor_n \times temperature\ efficiency_n$$ | **18** |
| Ventilation system | vent_inlet_temperature_code | nominal | Categorisation of ventilation heat recovery and heating coil, based on the maximum vent_mech_winter for the first three categories. If vent_mech_winter is zero, "Type 4" is selected.<br>• Type 1 = ventilation systems with temperature-controlled heat recovery (and temperature-controlled heating coil)<br>• Type 2 = ventilation systems with NOT temperature-controlled heat recovery and temperature-controlled heating coil<br>• Type 3 = ventilation systems with NOT temperature-controlled heat recovery and NO (temperature-controlled) heating coil<br>• Type 4 = no mechanical ventilation system | |
| Ventilation summer | vent_nat_summer | l/s | Total natural ventilation in summer calculated as stated in Equation 17. | |
| | vent_mech_summer | l/s | Total mechanical ventilation in summer calculated as: $$\sum_{n=1}^{i} area_n \times ventilation\ flow\ per\ area_n \times\ usage\ factor_n$$ | **19** |
| Solar plant | solar_plant_type_code | nominal | Type of solar plant:<br>• None = No solar plant (respectively solar plant with 0m$^2$ area)<br>• UtilityWater = only for domestic hot water<br>• RoomHeating = only for room heating<br>• Combined = Combined for room heating and domestic hot water | |
| | solar_plant_area | m$^2$ | Area of the solar plant | |

| | | | |
|---|---|---|---|
| Heat pump | heatpump_type_code | nominal | Type of solar plant:<br>• None = No heat pump (respectively heat pump with 0 area fraction)<br>• RoomHeating = only for room heating<br>• UtilityWater = only for domestic hot water<br>• Combined = One heat pump combined for room heating and domestic hot water<br>• Duo = Two heat pumps, one for room heating and one for domestic hot water |
| | heatpump_area_fraction | - | Proportion of the total heated floor area of the building covered by the heat pump. If heat pumps supply heat to the ventilation system's supply air, a negative number indicates that there is also other heating in the rooms. |

# Appendix A Quality criteria for EPC data

In Table 5 following criteria used for cleaning of the EPC data are outlined. These criteria are based on the work by [2,7] but were adapted, and some were added to fit the purpose of this dataset better. The column modified indicated if that criterion was compared to the ones used by [2,7] either modified or added.

**Table 5 EPC cleaning criteria based on the ones established by** [2,7]**. The column modified indicated if that criterion was compared to the one used by** [2,7] **either modified or added.**

| Component | Characteristic | Criterion | Modified |
|---|---|---|---|
| Building information | Heat capacity | [23,180] | ✓ |
| Building envelope information | Area | >0 | |
| | U-value | ]0.03, 7] | |
| | Temperature factor for roofs and ceilings | [0, 1] | ✓ |
| | Temperature factor external walls and floors | [0, 1.3] | ✓ |
| Window information | Number | >0 | |
| | Area | >0 | |
| | U-value | ]0.2, 7] | |
| | Temperature factor | [0, 1] | |
| | Fraction of glazing | [0, 1] | |
| | Solar transmittance (g-value) | [0, 1] | |
| | Shading angle horizon | [0, 90] | |
| | Shading angle eaves | [0, 90] | |
| | Shading angle left | [0, 90] | |
| | Shading angle right | [0, 90] | |
| Linear thermal transmittance information | Length | >0 | |
| | Heat loss | ]-0.1, 10] | ✓ |
| | Temperature factor | [0,1.3] | ✓ |
| Ventilation information | Area | >0 | |
| | Time of operation | [0, 1] | |
| | Natural ventilation winter | >=0 | |
| | Mechanical ventilation winter | >=0 | |
| | Natural ventilation summer | >=0 | ✓ |
| Ventilation information | Mechanical ventilation summer | >=0 | ✓ |
| | Heat recovery | [0, 1] | |
| | Inlet temperature | (-18, 0, 18) | |

| | | | |
|---|---|---|---|
| Internal heat gains | Area | >0 | |
| | Heat load from persons | ]0, 10] | |
| | Heat load from appliances inside and outside usage hours | ]0, 16] | |
| Heat distribution system | Supply temperature | [30, 90] | |
| | Return temperature | [15, 90] | |
| | Supply temperature & Return temperature | Exists at least once | ✓ |
| | Temperature difference | >=5 | ✓ |
| Heat/DHW distribution pipes | Length | >0 if temperature factor ≠ 0 | X |
| | Heat loss | >0 if temperature factor ≠ 0 | X |
| | Temperature factor | [0, 1] | |
| Domestic hot water | Average consumption | [0, 300] | |
| | Temperature | >=55 | ✓ |
| Domestic hot water tanks | Number | >= 0 | |
| | Volume | >= 0 | |
| | Share of DHW | [0, 1] | |
| | Supply temperature | [30, 90] | |
| | Heat loss | >0 | |
| | Temperature factor | [0, 1] | |
| Solar heating plant | Area | >= 0 | ✓ |
| Heat pump | Fraction of area | [-1, 1] | ✓ |