BuiltHub

Draft DELIVERABLE D3.2: Methodology on quality assurance

Project acronym	BuiltHub
Full title	Deliverable 3.2: Methodology on quality assurance
GA no	957026
WP, Deliverable #	3, D3.2
Version	1
Date	25.02.2022
Dissemination Level	Public
Deliverable lead	EURAC
Author(s)	Andrea Zambito, EURAC Simon Pezzutto, EURAC Ulrich Filippi Oberegger, EURAC
Reviewer(s)	David Lucas Reche, EVERIS Mikael Mangold, RISE
Keywords	Quality control, datasets, data, metadata, data comparison, data quality, data reliability, quality assurance

This project has received funding from the EU's Horizon 2020 programme under grant agreement no 957026.

Disclaimer

Table of contents

List of abbreviations

T	Task
D	Deliverable
WP	Work Package
GA	Grant Agreement
QC	Quality Control
EU	European Union
EC	European Commission
MS	Member State

1. Executive summary

The development of Task 3.2 (T3.2) is subsequential to the completion of Task 3.1 (T3.1) of the Horizon 2020 (H2020) project BuiltHub, Work Package 3 (WP3) – Data assembly. The aim is to report the datasets analyzed in T3.1 and apply a quality control (QC) process for guaranteeing high quality data provision. More information about the tasks can be found also in the Grant Agreement (GA) of given project [1].

The scope of Deliverable D3.1 [2] was to provide an overview on the status quo of the Building Stock Observatory (BSO) [44] database and to select the most relevant indicators which could have high importance and influence for future European policy makers. In relation to T3.1 of WP3 also a list of 30 datasets which potentially entail relevant data for filling the chosen indicators has been provided. This work brought to light several issues, among which the difficulty to find specific data, the presence of a high amount of data in an unstructured format and the presence of datasets reporting the same indicator but presenting different final values.

As stated by the Grant Agreement [1], when multiple options per single indicator are available, the services offered by BuiltHub will provide all possible options, giving to the final user the possibility to choose which data to select. Thus, it is necessary, for providing a high-level service, to have a quality control on the provided datasets. This is exactly the aim of T3.2 of WP3: A complete set of metadata needs to be provided for a correct and precise identification of the data sources and their quality level needs to be reported in the BuiltHub platform. Furthermore, different levels of quality control have been implemented, according to the GA More details concerning the metadata collection and quality control process can be found in the next chapters.

This deliverable (D3.2) - Methodology on quality assurance – entails information related to the metadata collection and to the quality control processes applied, aiming to increase the reliability of the provided datasets. This process should be a sign and indicator for high quality and high transparency of the datasets collected and included in the BuiltHub platform. This high data reliability will be one of the added values provided by BuiltHub, which will so appear to the users as a trustful and transparent platform. The quality control process will contribute to increase the confidence by the final users and also to make the collected and elaborated data more findable, accessible, interoperable, and re-usable (FAIR) [3].

The quality control is performed on the 30 datasets provided by D3.1 [2] at three different levels (more information is provided in the next chapters) but can serve as an example on how further datasets could be implemented in the platform in a second moment, guaranteeing so a continuous quality control of the provided data.

2. Introduction

Nowadays, among the most relevant threats there are climate change and natural resources depletion, leading to not only environmental but also social and economic issues [4]. Aware of these issues, policy makers are trying to push towards a more sustainable development of our economies [5]. A good example of this trend is the European Green Deal [6]. Through this action plan carried out by all Member States (MSs), Europe is trying to transform the Union in a modern, resource-efficient, and competitive economy. Policy making is supported by the gathering of data, which represents a milestone for a correct evaluation of the current situation and current trends. For this reason, in the last years several projects have been launched, aiming to collect and elaborate data from all European Community MSs (but also associated countries). The H2020 BuiltHub project is one of them. It focuses on the features and performances of the buildings, allowing the collection of important data and knowledge having a potential positive disruptive effect on the design by the European Commission (EC) of effective policies targeting buildings in view of the 2050 strategy and the European Green Deal [7].

The reasons for focusing on the building stock are multiple: First of all, in the EU buildings are responsible for about 40% of its energy consumption and 36% of greenhouse gas emissions, thus being a limit in reaching the carbon-neutrality. Furthermore, in Europe around 75% of the buildings can be considered inefficient, resulting in high energy losses. Renovating building stock could represent a significant action in the direction of reducing the emissions and in reaching the decarbonization target of the European Green Deal [8].

One of the most important projects for monitoring the European building stock has been launched in 2016 and is named “EU Building Stock Observatory (BSO)”. However, as already described in D3.1 the BSO presents several limitations and data lacks reducing its efficacy. For more information concerning the BSO data lacks and limitation please see D3.1 [2].

The aim of the BuiltHub project is not only a well-structured collection approach but also a benefits-based engagement strategy targeted to data and metadata providers, so to guarantee a continuous flow of building stock related data and metadata [7]. The added values provided by BuiltHub will be the basis for increasing the collaboration among BuiltHub itself and data owners, who will be so more convinced to feed-in data to the BuiltHub project, trying so to enlarge as much as possible the community related to the project itself guaranteeing a continuous and always larger data flow.

Deliverable D3.1 [2] reports a complete analysis of the BSO and of its limitations. It also provides a new selection of indicators chosen by the BuiltHub Consortium, which considered to be the most interesting for future policy making at a European level. Finally, a list of 30 datasets entailing relevant data for filling the chosen indicators has been provided. The Datasets entailed in this list are mainly of three types: building stock related datasets, socio-economic datasets, and climate-related datasets. These 30 datasets are the basis for the development of BuiltHub’s Task 3.2, and respective results are reported in this deliverable (D3.2).

The datasets collected not always report unique indicators, it happened in fact that a dataset could entail a specific indicator already reported by a second reported dataset. This leads to the overlapping of different data for the same indicator, but as reported by the GA: “the BuiltHub platform will not select one data among others - those will be shown equally side by side” [1]. On the other hand, BuiltHub aims to provide only reliable and trustful data to the final users,

which means that the datasets reported should overcome a quality control process. This quality process is divided in two main parts:

- **Metadata collection:** the first phase of the quality control phase consists in the collection of a series of metadata, which allows the user to clearly identify the source of the consulted data. More information concerning this phase is described in chapter 5. Metadata collection.

- **Quality control process:** this process is divided in 3 different quality levels, which aim to evaluate different statistical aspects of datasets presenting the same indicators. More details are reported in the introductive Chapter 4. Quality control levels and metadata collection. More specific information concerning the three different quality control levels are reported in Chapters 6. Datasets coming from stakeholder's consultation – Quality control level 1, 7. Consistency analysis – Quality control level 2, and 8. Comparison of similar datasets – Quality control level 3.

The quality control process is a fundamental step in the provision of reliable data to the final users. The scope of this process is described in the next Chapter 3. Scope. Please not that the results for the quality control process shown in chapters 7. Consistency analysis – Quality control level 2, 8. Comparison of similar datasets – Quality control level 3, and 10. Annex A, are all cases in which no big differences between the compared indicators have been found. In Chapter 11. Annex B, it is instead possible to find some examples of cases in which the difference between the compared indicators was relevant.

3. Scope

As already introduced in Chapter 2. Introduction there was an overlapping in the indicators provided by the 30 different datasets collected in T3.1 and reported in D3.1 [2]. In order to maintain all data collected leaving to the final users the decision of which data to use, it has been necessary to implement a quality control process for guaranteeing the provision of only reliable data.

The first step of the quality control, consisting in the metadata collection, is necessary for giving the possibility to final users to identify the sources for the data they are using. This means not only indicating the author and the link to the data, but also a number of metadata related to data availability, granularity, methodology for their gathering, terms of use and a lot more (for more information, please see Chapter 5. Metadata collection). This high data reliability will be one of the added values provided by BuiltHub, increasing the trustfulness and transparency of the platform itself. The quality control process will contribute to increase the confidence by the final users and also to make the collected and elaborated data more findable, accessible, interoperable, and re-usable (FAIR) [3].

The second step of the quality control process, as described in Chapter 4. Quality control levels and metadata collection, is split in 3 different quality control levels, where Level 1 quality control only entails collected datasets, which have been indicated in stakeholders' dialogues (T2.3); Level 2 includes assembled datasets, which underwent a consistency analysis control and level 3 contains solely datasets, which have been analyzed, as well as compared with related datasets (statistical comparison). In this way it is possible to guarantee to provide an indication about datasets having data out of range or same indicators with data having totally different behaviors. More details concerning the three quality control levels are given in the introductive Chapter 4. Quality control levels and metadata collection. More specific information concerning the three different quality control levels are reported in chapters 6. Datasets coming from stakeholder's consultation – Quality control level 1, 7. Consistency analysis – Quality control level 2, and 8. Comparison of similar datasets – Quality control level 3.

The quality control process implemented and reported in this deliverable aims to verify the quality of the 30 datasets collected in T3.1 and reported in D3.1 [2]. However, the methods described in this report are also useful for the future implementation of further datasets, so to know which metadata should be added and how to eventually compare the provided data with the already implemented ones.

4. Quality control levels and metadata collection

As already explained in Chapter 2. Introduction, the quality control process integrated in Task 3.2 and implemented on the 30 datasets collected in Task 3.1, basically consists in two main phases: the first is the metadata collection (see Subchapter 4.1. Metadata collection) and the second is a quality control process subdivided in three different levels and based on the evaluation of statistical parameters (see Subchapter 4.2. Quality control levels). The aforementioned subchapters aim to clarify which metadata have to be collected and the reasons why to collect them, but also the meaning of the three quality control levels and how they have to be implemented.

4.1. Metadata collection

One of the most important aspects to take into account when consulting a source or when using a certain data is their reliability. Data coming from trustful sources, explaining and demonstrating for example how the data has been gathered, are more reliable than others. One of the objectives of BuiltHub is to only provide reliable data, and in general to always provide to the final users the possibility to consult the set of metadata collected and decide so if to use one data or another provided by the platform. As already explained in deliverable D3.1 [2], two different metadata schemes have been taking as a basis for the metadata collection: DataCite scheme [9] and Schema.org scheme [10]. The metadata required by DataCite are more specifically author(s), title, DOI, publisher and publication year, and finally resource type. Among them it has been decided to avoid the collection only of the publisher, which is considered as a marginal information. Providing metadata in line with DataCite is an advantage since they will be coherent with the needs of large scientific repositories such as Zenodo. The metadata provided by DataCite scheme present a lack in metadata collection concerning semantic data related to the identified sources of information. For this reason, a second metadata scheme has been introduced: schema.org [10]. Here it follows a short list of the metadata selected for the metadata collection scheme, a full description of them can be found in Chapter 5. Metadata collection:

- Name
- Content
- Author/s
- Dataset URL
- Reference and publication year
- Spatial extension
- Granularity
- Methodology URL
- Methodology description
- Accuracy
- Completeness
- Source
- Access
- License
- Terms of Use
- Source type

4.2. Quality control levels

The second phase of the quality control process is split in three different levels according to the GA [1]. Among the datasets provided by the D3.1 [2], five of them had to fulfill the requirements of QC level 1, 10 of them the requirements of QC level 2 and the remaining 15 the ones of QC level 3. Metadata had instead to be provided for all 30 datasets implemented in D3.1. The aforementioned goals set by the GA are briefly reported in the following Table 1:

Table 1: Quality control levels - number of datasets to be provided at each quality control level [1]

Quality control level	Number of datasets
Quality control level 1	5/30
Quality control level 2	10/30
Quality control level 3	15/30

More information concerning the specific datasets and their specific QC level can be found in Table 2 (please note that the enumeration and the names used in the Table above are the same of the ones provided in the tables of D3.1 [2]). The complete results and methodologies for the three QC levels can be found in chapters 6. Datasets coming from stakeholder's consultation – Quality control level 1, 7. Consistency analysis – Quality control level 2, and 8. Comparison of similar datasets – Quality control level 3, while a brief introduction to them can be found in the following paragraphs:

- **Quality control level 1**

The QC level 1 only entails collected datasets, which have been indicated in stakeholders' dialogues (T2.3). Task 2.3 of the BuiltHub project aims in fact on the creation of a community of stakeholders, which should be the basis for the continuous feedback on the platform created: especially its user-friendliness, data availability, and services. The main instrument for the collection of stakeholders feedbacks are online surveys. The continuous dialogue between the stakeholders and the main target users of the platform is the main objective of Sub-Task 2.3, so to ensure the needs of the users of the platform itself. This continuous dialogue will further develop the relationship between BuiltHub and its community, which should improve the process of data gathering by encouraging direct involvement in data provision. Quality control level 1 is the simplest and simply guarantees the provision of a dataset which have been specifically asked during the stakeholders' dialogues. In Chapter 6. Datasets coming from stakeholder's consultation – Quality control level 1 it is possible to find the specific datasets among the 30 reported in D3.1 [2], which have been reported as quality control level 1. Only 5 datasets among the 30 of D3.1 had to be provided at a QC level 1.

- **Quality control level 2**

The QC level 2 is not related to the requirements of the stakeholders but consists in a consistency analysis of the considered datasets. It includes assembled datasets, which underwent the above-described quality control, but has not been compared to related datasets. More specifically, the consistency analysis is based on the evaluation of the p-value. Basically, if the p-value is below a pre-determined significance level (in our case the value of 5% has been chosen) then it is possible to state that there is a correlation. It does not exactly tell the extent of correlation, but just that there is evidence that there is correlation which for the purpose of QC level 2 is enough (if the variances of two datasets that should be consistent are correlated then it is possible to state that this supports that they are consistent, if one dataset was inconsistent then it would not be correlated with the other dataset). Ten datasets among the 30 of D3.1 have been provided at a QC level 2 [1], [2]. More information concerning the specific datasets chosen and the details concerning the p-value analysis can be found in Chapter 7. Consistency analysis – Quality control level 2. A number of cases of indicators having a p-value higher than 5% are reported in Chapter 11. Annex B.

- **Quality control level 3**

The QC level 3 is the most complete one. It contains all the information which have to be gathered according to the metadata provision, the p-value analysis performed for the QC level 2, but also more specific statistical analysis. The QC level 3 contains solely datasets, which have been analyzed, as well as compared with related datasets. The aim is to compare the same indicator provided by different datasets among the ones selected in D3.1 [2], understanding if the trend of the data and their values distribution are similar. In order to achieve this purpose, the following parameters are calculated for each dataset:

- Minimum and maximum values of the dataset
- Standard deviation
- 25^{th}, 50^{th} and 75^{th} percentiles (first, median and third quartiles)

These statistical parameters are compared for both the selected datasets and a minimum percentage difference is set, so to guarantee an indication about if both the absolute values and the general trend and distribution of the data are similar, underlining so the case when percentage differences are higher than 30%. More information concerning QC level 3 and the specific datasets evaluated as QC level 3 is given in Chapter 8. Comparison of similar datasets – Quality control level 3. Fifteen datasets among the 30 of D3.1 have been provided at a QC level 3 [1], [2].

A complete list of the selected datasets in D3.1 is reported in the following Table 2, where also the quality control level chosen and applied to each dataset is reported. More complete data concerning the specific quality control results can be found in the next chapters 6. Datasets coming from stakeholder's consultation – Quality control level 1, 7. Consistency analysis – Quality control level 2, and 8. Comparison of similar datasets – Quality control level 3.

Table 2: List of datasets provided by the Deliverable 3.1 of BuiltHub [2], with the addition of the proved quality control level and identification of eventual comparison performed (related dataset chosen and specific indicator compared)

Quality control (QC) number:	Dataset N.1	Dataset N.2	Indicator compared	QC level 1	QC level 2	QC level 3
1	Horizon 2020 HotMaps project: Building stock analysis [11]	FP7 CommONEnergy Project: building stock [12]	Buildings Floor area: non-residential sector [Mm^2]		X	
2	IEE TABULA project: Typology Approach for Building Stock Energy Assessment [13]	Horizon 2020 HotMaps project: Building stock analysis [11]	Thermal transmittance of building components: U-value [$W/(m^2K)$]			X
3	IEE EPISCOPE project: Focus of building stock monitoring [14]	Horizon 2020 HotMaps project: Building stock analysis [11]	Buildings Floor area: residential sector [Mm^2]		X	
4	IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020 [15]	Horizon 2020 HotMaps project: Building stock analysis [11]	Thermal transmittance of building walls (2010-2016): U-value [$W/(m^2K)$]			X
5	IEE ENTRANZE project: Policies to Enforce the TRAnsition to Nearly Zero Energy buildings in the EU27 [16]	National Housing Census: European statistical System [17]	mean single dwelling surface [m^2]			X
6	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland. [18]	Horizon 2020 HotMaps project: Building stock analysis [11]	Buildings Floor area: residential sector [Mm^2]			X

7	FP7 CommONEnergy Project: building stock [12]	IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020 [15]	Non-residential sector - built floor area [m^2]		X	
8	JRC IDEES 2015 [19]	Horizon 2020 HotMaps project: Building stock analysis [11]	Final energy consumption for space heating in residential sector [Mtoe]		X	
9	SET-Nav - Strategic Energy Roadmap [20]	EUROSTAT: Final energy consumption in households by fuel [21]	Final energy consumption in households [Mtoe]		X	
10	H2020 ExcEED Project: building stock data [22]	-	-	X		
11	FP7 iNSPiRe project: building stock analysis [23]	JRC IDEES 2015 [19]	Final energy consumption for lighting in residential sector [TWh/year]		X	
12	Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT [24]	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia, and Switzerland. [18]	Electricity consumption in households in 2016 [GWh/year]			X
13	Comprehensive study of building energy renovation activities and the uptake of nearly zero-energy buildings in the EU - FINAL REPORT [25]	EUROSTAT: Population on 1 January by age, sex, and NUTS 2 region [26]	Population by country in 2018 [person]			X
14	EUROSTAT: Final energy consumption in households [27]	JRC IDEES 2015 [19]	Final energy consumption in households in 2015 [kToe]			X

15	EUROSTAT: Final energy consumption in households by fuel [21]	JRC IDEES 2015 [19]	Final energy consumption - share by fuel - solids fossil fuels [%]		X	
16	EUROSTAT: Disaggregated final energy consumption in households [28]	Horizon 2020 HotMaps project: Building stock analysis [11]	Final energy consumption in households [kToe]		X	
17	ZENSUS 2011 [29]	National Housing Census: European statistical System [17]	Number of dwellings by size of private household in Germany at NUTS2 level in year 2011			X
18	DPE - Diagnostic de Performance Energetique [30]	-	-	X		
19	BPIE - EUROPE'S BUILDINGS UNDER THE MICROSCOPE [31]	Horizon 2020 HotMaps project: Building stock analysis [11]	Share of owner-occupied dwellings in the residential sector [%]			X
20	DEEP - De-risking Energy Efficiency Platform [32]	-	-	X		
21	European Union energy statistical pocketbook -2019 update [33]	EDGAR (Emissions Database for Global Atmospheric Research) CO_2 Emissions [34]	Tons of carbon dioxide emissions related to the residential sector [tCO_2]			X
22	Dataset of the publication: Europe's Building Stock and Its Energy Demand: A Comparison Between Austria and Italy [35]	Horizon 2020 HotMaps project: Building stock analysis [11]	Residential and Offices shares in constructed square meters per construction vintage - Italy and Austria [%]		X	
23	National Housing Census: European statistical System [17]	Horizon 2020 HotMaps project: Building stock analysis [11]	Number of occupied dwellings in residential sector [mil]			X

24	Energy prices in 2019 - Household energy prices in the EU [36]	Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT [24]	Electricity prices [EUR/kWh]			X
25	EUROSTAT: GDP per capita in PPS [37]	WorldBank: GDP in PPP [38]	GDP in PPS for the European countries			X
26	EUROSTAT: Population on 1 January by age, sex and NUTS 2 region [26]	National Housing Census: European statistical System [17]	Population at NUTS2 level [person]			X
27	EUROSTAT - Cooling and heating degree days [39]	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland. [18]	Heating degree days (HDD)			X
28	EDGAR (Emissions Database for Global Atmospheric Research) CO_2 Emissions [34]	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland. [18]	Tons of carbon dioxide emissions related to the residential sector [tCO_2]		X	
29	CORDEX - Regional climate model data on single levels for Europe [40]	-	-	X		
30	PVGIS - Photovoltaic Geographical Information System [41]	-	-	X		

*Please note that datasets number 19 and number 20 selected in D3.1 have been substituted by the following two datasets, given the impossibility to access the data or given the provision of only disaggregated data referring to years before 2005. The selected substitutes are: 19. BPIE - Europe's buildings under the microscope; Marina Economidou et Al. (BPIE). 2011. https://bpie.eu/wp-content/uploads/2015/10/HR_EU_B_under_microscope_study.pdf [31] and 21. European Union energy statistical pocketbook. DG Energy. 2019. https://data.europa.eu/data/datasets/eu-energy-statistical-pocketbook?locale=en [33]

5. Metadata collection

As anticipated in Sub-Chapter 4.1 Metadata collection, the provision of metadata related to the provided data increases the reliability of the data themselves. Metadata are fundamental for the provision of different information to the final user, among them there are information related to:

- Who is related to the data source (authors);
- What is the source about (name, content, reference year, granularity, spatial extension);
- Where to find this source (dataset URL, publication year);
- How data have been gathered (methodology, methodology URL);
- Quality of the dataset (accuracy, completeness, source);
- Accessibility to the data (access, license, terms of use, source type)

As already introduced in Sub-chapter 4.1 Metadata collection, two different metadata schemes have been taking as a basis for the metadata collection: DataCite scheme [9] and Schema.org scheme [10]. DataCite metadata are mostly focused on the collection of data related to authors, contents, and accessibility to the source (information about publishers and similar), but mostly miss metadata collection concerning semantic data. For this reason, DataCite scheme has been integrated with the schema.org scheme [10].

The following Table 3 shows the complete list of metadata chosen by the BuiltHub Consortium that will be connected to the data the BuiltHub platform is going to provide. A complete description of all the chosen metadata is provided as well:

Table 3: Metadata list and complete description

MATADATA	INFORMATION CONTAINED
NAME	Is the complete name of the source (e.g. datasets, document, report) in which the collected information is reported.
CONTENT	It is a brief description of the content of the dataset or report taken into consideration. Gives to the reader an idea of the main data and information she/he can find in the document itself.
AUTHOR/S	Both the organization and/or individuals responsible for producing the data are reported as authors.
DATASET URL	It is a unique, persistent code or link that can be used to locate the dataset for collection. Generally, the dataset

	URL is reported as an internet address directly linking to the identified source.
REFERENCE YEAR	It is the year/s for which data are provided in the selected data source. It can be a single year, a mean value for a certain time frame or multiple years.
PUBLICATION YEAR	It is intended as the year that the data was published.
SPATIAL EXTENSION	As spatial extension it is intended the whole covered area by the dataset (e.g. EU27+UK, World, Europe, Single states/regions).
GRANULARITY	It is intended as the smallest spatial unit that the dataset can cover (e.g. NUTS0, NUTS2, LAU1, 100x100m grid).
METHODOLOGY URL	It is a specific internet address linking to a more detailed description of the methodology used by authors for gathering/elaborating the data contained in the indicated data source.
METHODOLOGY DESCRIPTION	It is a brief description of the methodology used by the authors for gathering/elaborating the data contained in the indicated data source.
ACCURACY	It provides an indication of the accuracy of the reported data by the selected dataset/source. It is usually provided by the author/s itself.
COMPLETENESS	It gives an idea of how much complete the considered dataset is, according to its reference year/s, spatial extension, and granularity. This information is often provided by the author/s.
SOURCE	As “source” it is intended the source effort of the data (for example, the project that produced the dataset).
ACCESS	It entails a description of whether the data is open and available for download, if the data are confidential or even if a registration is required for accessing the data.
LICENSE	It refers to the license detailing the use conditions for the data reported.
TERMS OF USE	Any brief details on the terms of use for the data.
SOURCE TYPE	It intends the type of object reported (e.g. dataset, PDF report, Book section).

*The metadata mentioned in this table have been provided, whenever possible, for all the 30 datasets collected in relation to Sub-task 3.1 and have been party reported in D3.1 [2].

6. Datasets coming from stakeholder's consultation – Quality control level 1

Quality control level 1 is the first and mostly connected to the BuiltHub stakeholder's quality assurance level. In fact, it only entails collected datasets, which have been indicated in stakeholders' dialogues (T2.3). One of the most relevant objectives of the BuiltHub project is to ensure a continuous flow of buildings-related data creating around the project itself a strongly connected community of users/stakeholders providing and at the same time using data. The community of stakeholders are the basis also for the continuous feedback on the platform created. The main feedbacks provided are in terms of user-friendliness, availability of data, and services. In order to communicate with stakeholders and to collect their feedbacks the main instrument utilized by BuiltHub are online surveys. The continuous dialogue with stakeholders guarantees that the need of final users will be met by the platform itself. Quality control level 1 is the simplest and simply guarantees the provision of a dataset which have been specifically asked during the stakeholders' dialogues, The GA itself, as also reported in Table 1, required to provide 5 datasets among the 30 selected in Sub-task 3.1 (see D3.1 [2]) at QC level 1. The following Table 4 contains the complete list of datasets of D3.1 provided with a quality control of level 1.

Table 4: List of datasets of D3.1 [2] provided with a quality control of level 1. Please note that the number in the first column refers to the enumeration used in Deliverable 3.1 for the list of 30 datasets for data provision to the BuiltHub platform

Quality control (QC) number:	Dataset Name	QC level 1
10	H2020 ExcEED Project: building stock data [22]	X
18	DPE - Diagnostic de Performance Energetique [30]	X
20	DEEP - De-risking Energy Efficiency Platform [32]	X
29	CORDEX - Regional climate model data on single levels for Europe [40]	X
30	PVGIS - Photovoltaic Geographical Information System [41]	X

7. Consistency analysis – Quality control level 2

Quality control level 2 is based on the application of a consistency analysis on the considered datasets. The consistency analysis, as already introduced in Sub-chapter 4.2 Quality control levels, is based on the evaluation of the p-value through a simple linear regression, so to determine if there could be evidence of correlation between the two considered datasets. The methodology employed for the QC level 2 can be schematized in four main points:

1. **Selection of the related dataset**
   Consists in the selection of a dataset which entails the same data of the dataset we want to evaluate, or even a dataset linearly correlated with this one (for example the space heating consumption in residential sector and the population density in specific regions). The aim of this analysis is to understand if the two datasets have the same behaviour so to be sure not to provide data presenting not realistic information. This is important also considering that BuiltHub aims to provide, when available, multiple data for even single indicators. It is clear that this approach allows to be sure not to provide data for the same indicator presenting different behaviours.

2. **Selection of the subset data**
   Not all the datasets provided in D3.1 have the same granularity, spatial extension or even unit of measure. Thus, it is necessary to elaborate the dataset so to obtain a subset of data which can be analyzed in a one-be-one data approach with the related dataset. In the case of the analysis performed in relation to Sub-task 3.2 only datasets presenting the same indicator have been used. In the case of presence of datasets being the only among the 30 provided to report a certain indicator, external sources have been selected and used for the analysis. In this case more information concerning the selected external sources are reported in Table 8.

3. **Implementation of the regression**
   Once the assessed data subset data and the related datasets have been collected, a linear regression is performed, imposing the first one as the independent variable and the second one as the dependent variable. The linear regression analysis provided all information shown in the example of Table 5, Table 6, and Table 7 [43]. The data we wanted to obtain was the p-value (Table 7), indicating if the correlation between the two considered variables (which in this specific case are representative for the two analyzed datasets) is statistically significant. It has been decided to use the significance level of 5% for identifying if the correlation was statistically significant (if the p-value is lower than 0.05). The Null hypothesis used for testing the significance of correlation is clearly that no correlation (linear relationship) between the two variables occurs. A p-value lower than 5% indicates so a rejection of the null hypothesis, showing that the possibility that the two considered datasets are not correlated is lower than 5%. Generally, the lower is the p-value, the higher is the probability that the two datasets considered are consistent [42].

4. **Quality control acceptance**
   Once the linear regression is implemented and the p-value is calculated, the dataset quality can be evaluated according to the distance from the reference value of 5% imposed.

The datasets required at a quality control level 2 check are 10 (for further information, please see Table 1). QC level 2 consistency analysis is the basis also for the QC level 3, in which further statistical analysis are performed.

Table 5: Statistical parameters provided by the linear regression performed using Excel considering two comparable datasets - regression statistics [43]

Regression Statistics	
Multiple R *(1)	0.98
R Square *(2)	0.97
Adjusted R Square *(3)	0.97
Standard Error *(4)	1.15
Observations *(5)	29

***(1) Multiple R:** correlation coefficient. It shows how strong the linear relationship is. For example, a value of 1 means a perfect positive relationship and a value of zero means no relationship at all. It is the square root of R square (please see *(2)).
***(2) R squared:** or r^2, is called the Coefficient of Determination. It indicates how many points fall on the regression line. For example, 80% means that 80% of the variation of y-values around the mean are explained by the x-values. In other words, 80% of the values fit the model. The higher this value the higher is the compatibility between the two selected datasets.
***(3) Adjusted R square:** The adjusted R-square adjusts for the number of terms in a model. This parameter can be used instead of *(2) in the case of presence of more than a single variable x, but this is not the case of our study.
***(4) Standard Error of the regression:** An estimate of the standard deviation of the error μ. The standard error of the regression is the precision that the regression coefficient is measured; in the case the coefficient is large if compared to the standard error, then the coefficient is probably different from 0.
***(5) Observations:** Number of observations in the sample. The bigger the sample the more precise will be the analysis.

Table 6: Statistical parameters provided by the linear regression performed using Excel considering two comparable datasets – ANOVA [43]

ANOVA					
	df	SS *(1)	MS	F	Significance F
Regression	1	1117.74	1117.72 *(2)	838.52 *(4)	7.22474E-22 *(5)
Residual	27	35.99	1.33 *(3)		
Total	28	1153.73			

This second part of the data analysis provided by the data analysis tool of Excel is generally less used since It splits the sum of squares into individual components.
*(1) SS = Sum of Squares.
*(2) Regression MS = Regression SS / Regression degrees of freedom.
*(3) Residual MS = mean squared error (Residual SS / Residual degrees of freedom).
*(4) F: Overall F test for the null hypothesis.
*(5) Significance F: The significance associated P-Value.

Table 7: Statistical parameters provided by the linear regression performed using Excel considering two comparable datasets - p-value [43]

	Coeff.	Std Error	t Stat.	p-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.118	0.234	-0.506	0.617	-0.598	0.3613	-0.598	0.361
JRC IDEES - 2015	0.907	0.031	28.96	7.22474E-22	0.843	0.971	0.843	0.971

*(1) Coefficient: Gives you the least squares estimate.
*(2) Standard Error: The least squares estimate of the standard error.
*(3) T Statistic: The T Statistic for the null hypothesis vs. the alternate hypothesis.
*(4) P Value: Gives you the p-value for the hypothesis test.
*(5) Lower 95%: The lower boundary for the confidence interval.
*(6) Upper 95%: The upper boundary for the confidence interval.

The following Table 8 reports the full list of datasets which underwent the quality control level 2. Moreover, also the intercept dataset and the indicator compared have been reported. All 15 datasets analyzed present a p-value lower than 5%, which means having a successful consistency analysis. The process described in the QC level 2 is the basis for performing the QC level 3 described in Chapter 8. Comparison of similar datasets – Quality control level 3. This means that all the datasets which passed QC level 3 automatically already showed to have a p-values lower than 5% even if this is not reported in the following resuming tables.

Table 8: List of assessed and related datasets pairs, comprehensive of name of the indicator considered in the regression analysis, quality control level performed, and final results gained. Please note that the quality control number refers to Table 2 and to Deliverable D3.1 [2].

Quality control (QC) number:	Dataset N.1	Dataset N.2	Indicator compared	QC level 2	p-value
1	Horizon 2020 HotMaps project: Building stock analysis [11]	FP7 CommONEnergy Project: building stock analysis [12]	Buildings Floor area: non-residential sector [Mm^2]	X	< 5%
3	IEE EPISCOPE project: Focus of building stock monitoring [14]	Horizon 2020 HotMaps project: Building stock analysis [11]	Buildings Floor area: residential sector [Mm^2]	X	< 5%
7	FP7 CommONEnergy Project: building stock analysis [12}	IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020 [15]	Non-residential sector - built floor area [m^2]	X	< 5%

8	JRC IDEES 2015 [19]	Horizon 2020 HotMaps project: Building stock analysis [11]	Final energy consumption for space heating in residential sector [Mtoe]	X	< 5%
9	SET-Nav - Strategic Energy Roadmap [20]	EUROSTAT: Final energy consumption in households by fuel [21]	Final energy consumption in households [Mtoe]	X	< 5%
11	FP7 iNSPiRe project: building stock analysis [23]	JRC IDEES 2015 [19]	Final energy consumption for lighting in residential sector [TWh/year]	X	< 5%
15	EUROSTAT: Final energy consumption in households by fuel [21]	JRC IDEES 2015 [19]	Final energy consumption - share by fuel - solids fossil fuels [%]	X	< 5%
16	EUROSTAT: Disaggregated final energy consumption in households [28]	Horizon 2020 HotMaps project: Building stock analysis [11]	Final energy consumption in households [kToe]	X	< 5%
22	Dataset of the publication: Europe's Building Stock and Its Energy Demand: A Comparison Between Austria and Italy [35]	Horizon 2020 HotMaps project: Building stock analysis [11]	Residential and Offices shares in constructed square meters per construction vintage - Italy and Austria [%]	X	< 5%
28	EDGAR (Emissions Database for Global Atmospheric Research) CO_2 Emissions [34]	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland [18]	Tons of carbon dioxide emissions related to the residential sector [tCO_2]	X	< 5%

8. Comparison of similar datasets – Quality control level 3

The quality control level 3 is the last and most complete QC level developed within the quality control process. As for QC of levels 1 and 2 a complete set of metadata has to be provided (for more information see Chapter 5. Metadata collection). In addition to the provision of metadata, the data analysis described in QC level 2 concerning the consistency analysis has to be performed, ensuring a p-value lower than 5% also for QC level 3. The difference between QC level 2 and QC level 3 is a more specific statistical analysis based on the comparison of datasets reporting the same indicators/data. According to the GA, the QC level 3 contains solely datasets, which have been analyzed, as well as compared with related datasets. The aim of QC level 3 is to compare the same indicator provided by different datasets among the ones selected in D3.1 [2], so to guarantee that the trend of the data and their values distribution are similar. The main statistical parameters evaluated for each dataset and used for the comparison with other datasets are listed in Table 9. These statistical indicators are the one describing the general behaviour and distribution of a set of data and are so useful for comparing related datasets.

Table 9: List of statistical indicators evaluated in Quality control level 3 and used for comparing related datasets

Statistical indicators for Quality control level 3	
Minimum value	Lowest value of the dataset
Maximum value	Highest value of the dataset
Standard deviation	Standard deviation is a statistic parameter measuring the dispersion of a dataset relative to its mean. It is calculated as the square root of variance by determining each data point's deviation relative to the mean. When the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.
25th percentile	Also called first, or lower, quartile. The 25th percentile is the value at which 25% of the collected data lie below that value, and 75% of them lie above that value.
50th percentile	Also known as the Median. The median cuts the data set in half. Half of the considered data lie below the median and half lie above the median.
75th percentile	Also defined as third, or upper, quartile. The 75th percentile is the value at which 25% of the considered data lie above that value and 75% of them lie below that value.

The aforementioned statistical parameters (Table 9) are collected for all analyzed datasets and compared for couples of datasets reporting the same indicator. As done for QC level 2, when comparing two different datasets, the same granularity, data availability and unit of measure had to be used, introducing so an elaboration process for all the datasets utilized.

Once obtained the subset to be compared and collected all the aforementioned statistical parameters a simple and automatic way for comparing the two results had to be found. It has been internally decided to simply set a threshold percentage difference between all the comparable statistical parameters in order to accept the quality control level 3 check. More in the detail, all pairs of statistical parameters (the two minimum values, the two standard deviations, etc.) have been compared and a percentage difference has been calculated. In the case of all 6 parameters difference was below the threshold value of 30%, it has been decided to confirm the compliance with the requirement of quality control level 3 (together with the compliances required by quality control level 2 and metadata provision). As required by the GA and showed in Table 1, 15 datasets at quality control level 3 have been provided. The complete list is reported in Table 10Table 9. The complete set of statistical parameters calculated for the 15 pairs of datasets is reported both in graphical and tabular form in Annex A. Table 10 entails the main results for 15 datasets which underwent QC level 3 showing all parameters with differences lower than 30%. Sometimes the differences where however higher, thus some examples are entailed in Annex B.

Table 10: List of datasets which underwent the quality control procedure of level 3. Please note that the fulfillment of the agreed factors has been reported for each pair of datasets. Please note that the quality control number in the first column is the same reported in the deliverable D3.1 [2] of BuiltHub and in Table 2. More information and the complete set of results can be found in Annex A.

Quality control (QC) number:	Dataset N.1	Dataset N.2	Indicator compared	Quality control (QC) level 3	p-value - Y (if <5%) or N (if >5%)	Difference in percentage between Dataset N.1 and dataset N.2 – [%]					
						Standard deviation	Minimum	25 percentiles	50 percentiles	75% percentiles	Maximum
2	IEE TABULA project: Typology Approach for Building Stock Energy Assessment [13]	Horizon 2020 HotMaps project: Building stock analysis [11]	Thermal transmittance of building components: U-value [W/(m^{2}K)]	X	Y	6.2	10.2	3.1	0.4	13.5	3.5

4	IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020 [15]	Horizon 2020 HotMaps project: Building stock analysis [11]	Thermal transmittance of building walls (2010-2016): U-value [W/(m^{2}K)]	X	Y	22.1	22.3	29.6	19.3	2.8	13.4
5	IEE ENTRANZE project: Policies to Enforce the TRAnsition to Nearly Zero Energy buildings in the EU27 [16]	National Housing Census: European statistical System [17]	mean single dwelling surface [m^2]	X	Y	25.4	23	0.8	5.2	7.7	8.1
6	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia, and Switzerland [18]	Horizon 2020 HotMaps project: Building stock analysis [11]	Buildings Floor area: residential sector [Mm^2]	X	Y	14.8	17.1	8.8	15.5	24.6	11.2
12	Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT [24]	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia, and Switzerland [18]	Electricity consumption in households in 2016 [GWh/year]	X	Y	0.4	0	0	0.1	0.2	1.1

13	Comprehensive study of building energy renovation activities and the uptake of nearly zero-energy buildings in the EU - FINAL REPORT [25]	EUROSTAT: Population on 1 January by age, sex, and NUTS 2 region [26]	Population by country in 2018 [person]	X	Y	0.1	1.8	0.3	0.1	0.1	0.1
14	EUROSTAT: Final energy consumption in households [27]	JRC IDEES 2015 [19]	Final energy consumption in households in 2015 [kToe]	X	Y	3.1	0	0.1	0	1.1	3.3
17	ZENSUS 2011 [29]	National Housing Census: European statistical System [17]	Number of dwellings by size of private household in Germany at NUTS2 level in year 2011	X	Y	1.5	3.6	0.9	0.1	0.8	0.9
19	BPIE - EUROPE'S BUILDINGS UNDER THE MICROSCOPE [31]	Horizon 2020 HotMaps project: Building stock analysis [11]	Share of owner-occupied dwellings in the residential sector [%]	X	Y	3.7	5.2	2.2	0	5	1
21	European Union energy statistical pocketbook - 2019 update [33]	EDGAR (Emissions Database for Global Atmospheric Research) CO_2 Emissions [34]	Tons of carbon dioxide emissions related to the residential sector [tCO_2]	X	Y	26.6	1.7	12.1	30	16.9	29.1
23	National Housing Census: European statistical System [17]	Horizon 2020 HotMaps project: Building stock analysis [11]	Number of occupied dwellings in residential sector [mil]	X	Y	15.2	4.5	6.4	20.3	16.1	16.2

24	Energy prices in 2019 - Household energy prices in the EU [36]	Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT [24]	Electricity prices [EUR/kWh]	X	Y	3.9	1	8.3	10.6	4.3	5.8
25	EUROSTAT: GDP per capita in PPS [37]	WorldBank: GDP in PPP [38]	GDP in PPS for the European countries	X	Y	0.8	0.2	0.6	0.3	0.3	0.2
26	EUROSTAT: Population on 1 January by age, sex, and NUTS 2 region [26]	National Housing Census: European statistical System [17]	Population at NUTS2 level [person]	X	Y	0.9	0	0.4	0.4	0.3	1.9
27	EUROSTAT - Cooling and heating degree days [39]	H2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia, and Switzerland [18]	Heating degree days (HDD)	X	Y	0.4	0.2	0.2	0.1	0.1	0.3

9. Conclusions

A clear method for defining and performing a quality control on the provided datasets has been reported in this deliverable. The provided solutions can be implemented also for future upcoming datasets that could so be added to the BuiltHub platform guaranteeing their good quality level. 5 datasets at QC level 1, 10 at QC level 2 and 15 at QC level 3 have been supplied. This procedure, as already explained in Chapter 2 Introduction, provides an added value to the final users, who will have at their disposal highly reliable data and metadata, so to be able to consciously choose which data they prefer to use.

10. Annex A

This annex entails the complete set of data evaluated and compared for quality control of level 3 both in a tabular and in a graphical representation. Please note that the enumeration of each pair of datasets compared refers to the list of 30 datasets provided by Deliverable 3.1 of the BuiltHub project. The enumeration is the same also of the one reported in Table 2.

Table 11: Quality control level 3 main statistical data - tabular representation - datasets comparison number 2 [13], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	IEE TABULA project - mean U-Value [W/(m²K)]	Horizon2020 HotMaps project - mean U-Value [W/(m²K)]
count	20	20
standard deviation	1.0	1.0
minimum	0.3	0.4
25%	0.80	0.78
50%	1.10	1.10
75%	2.44	2.11
maximum	4.6	4.4

Figure 1: Quality control level 3 main statistical data - graphical representation - datasets comparison number 2 [13], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

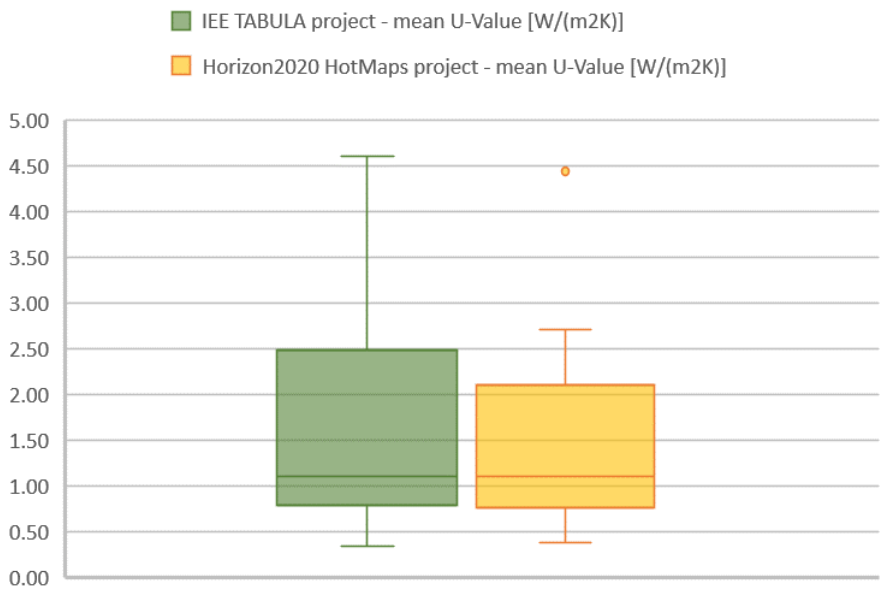


Table 12: Quality control level 3 main statistical data - tabular representation - datasets comparison number 4 [15], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	IEE Zebra2020 project - Mean U-value Walls 2010-2016 [W/m^{2}K]	Horizon 2020 HotMaps project - Mean U-value Walls 2010-2016 [W/m^{2}K]
count	16	16
standard deviation	0.2	0.2
minimum	0.2	0.2
25%	0.30	0.21
50%	0.32	0.30
75%	0.39	0.38
maximum	0.9	0.8

Figure 2: Quality control level 3 main statistical data - graphical representation - datasets comparison number 4 [15], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

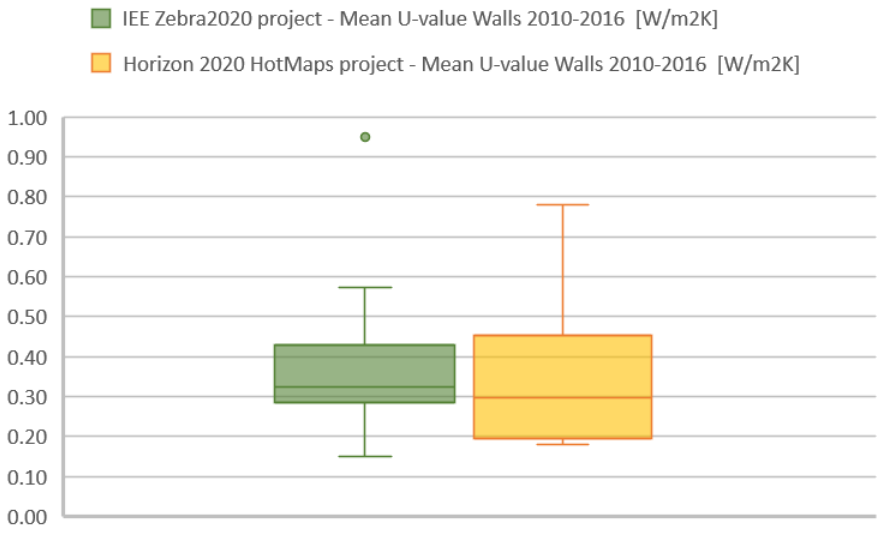


Table 13: Quality control level 3 main statistical data - tabular representation - datasets comparison number 5 [16], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	IEE Entranze project - mean dwelling surface [m^2]	National Census - mean dwelling surface [m^2]
count	25	25
standard deviation	16.6	22.2
minimum	57.0	43.9
25%	75.00	75.60
50%	84.00	88.60
75%	92.00	99.70
maximum	130.0	141.4

Figure 3: Quality control level 3 main statistical data - graphical representation - datasets comparison number 5 [16], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

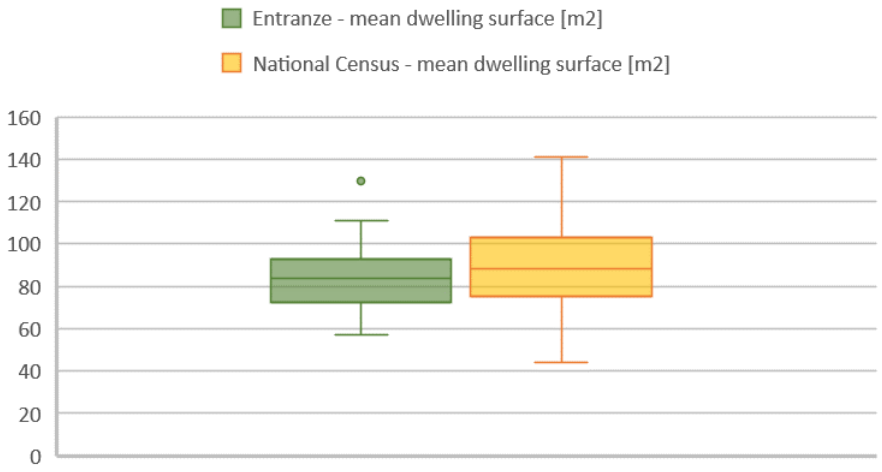


Table 14: Quality control level 3 main statistical data - tabular representation - datasets comparison number 6 [18], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Horizon2020 Odyssee project - Floor area residential sector 2016 [Mm^2]	Horizon2020 HotMaps project - Floor area residential sector 2016 [Mm^2]
count	22	22
standard deviation	1107.5	944.1
minimum	43.4	36.0
25%	154.11	140.59
50%	359.94	304.29
75%	833.76	628.93
maximum	3824.2	3395.6

Figure 4: Quality control level 3 main statistical data - graphical representation - datasets comparison number 6 [18], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

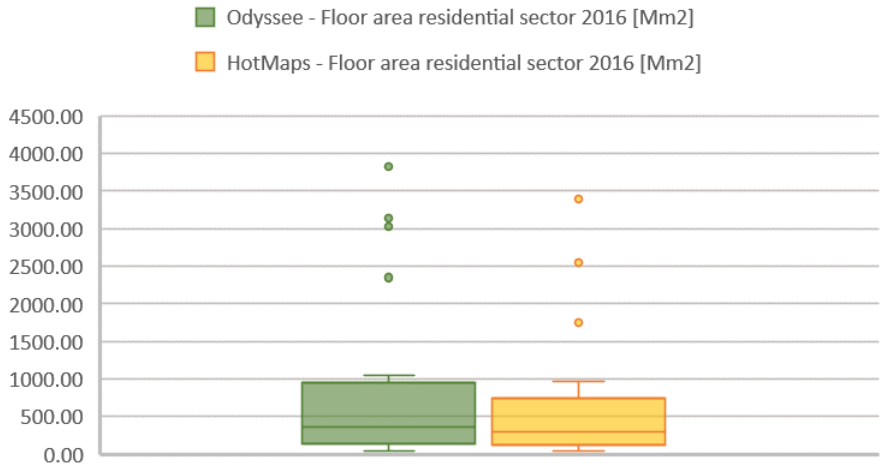


Table 15: Quality control level 3 main statistical data - tabular representation - datasets comparison number 12 [24], [18] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT - Electricity consumption in households (2016) - [GWh]	Horizon2020 Odyssee project - Electricity consumption in households (2016) - [GWh]
count	27	27
standard deviation	37762.8	37594.6
minimum	665.0	665.2
25%	4178.50	4178.66
50%	12067.00	12083.57
75%	22584.00	22549.41
maximum	159396.0	157663.3

Figure 5: Quality control level 3 main statistical data - graphical representation - datasets comparison number 12 [24], [18] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

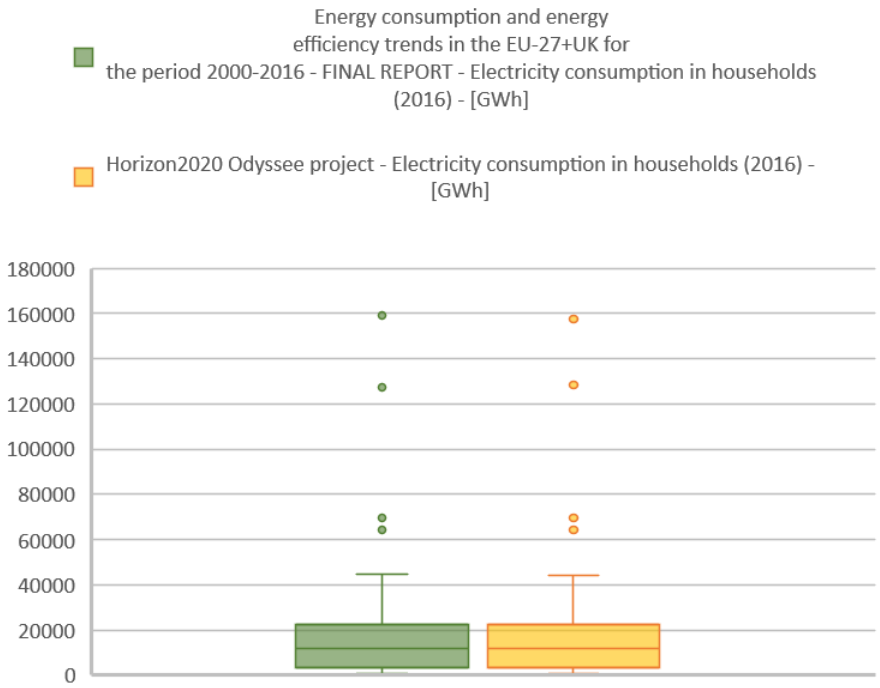


Table 16: Quality control level 3 main statistical data - tabular representation - datasets comparison number 13 [25], [26] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Comprehensive study of building energy renovation activities and the uptake of nearly zero-energy buildings in the EU - population 2018 [inhabitants]	EUROSTAT - population 2018 [inhabitants]
count	28	28
standard deviation	23334416.4	23354727.2
minimum	475701.0	484630.0
25%	3781345.00	3768538.25
50%	9300319.00	9308042.50
75%	17768470.75	17792210.50
maximum	82792351.0	82905782.0

Figure 6: Quality control level 3 main statistical data - graphical representation - datasets comparison number 13 [25], [26] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

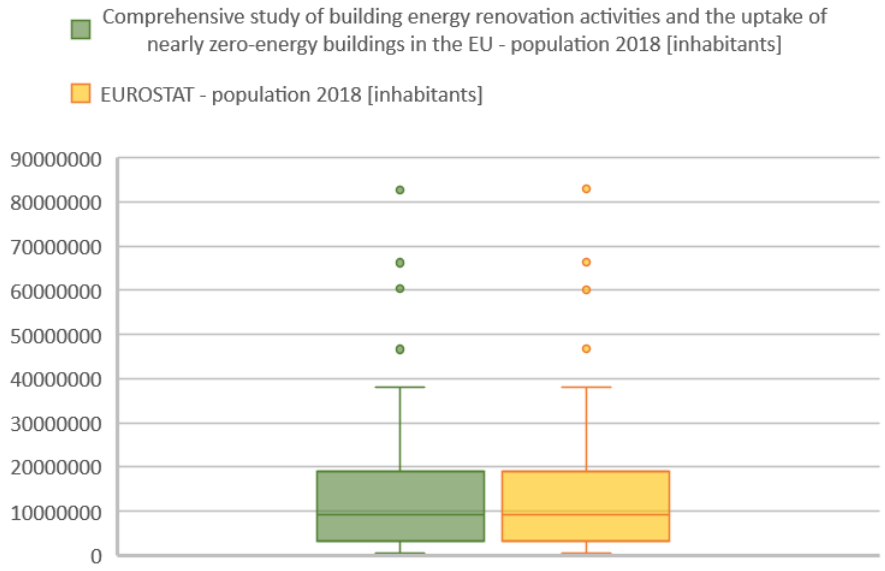


Table 17: Quality control level 3 main statistical data - tabular representation - datasets comparison number 14 [19], [27] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	JRC IDEES 2015 - Final energy consumption in households [kTOE]	EUROSTAT - Final energy consumption in households [kTOE]
count	28	28
standard deviation	13341.9	13775.7
minimum	0.0	0.0
25%	1832.04	1830.53
50%	4705.70	4705.80
75%	8498.02	8589.59
maximum	53171.0	54959.8

Figure 7: Quality control level 3 main statistical data - graphical representation - datasets comparison number 14 [19], [27] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

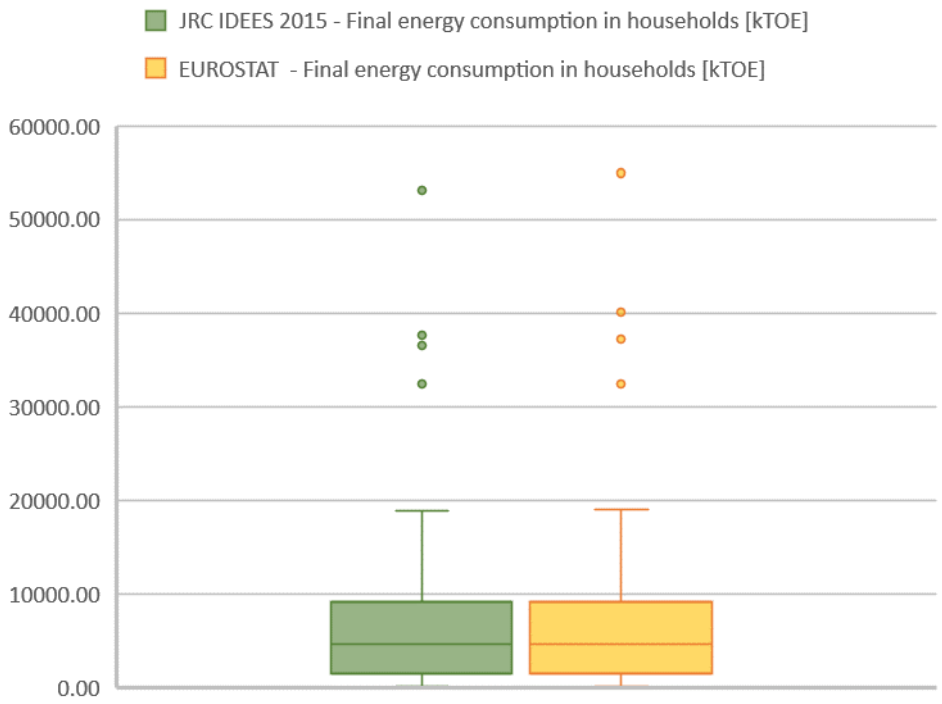


Table 18: Quality control level 3 main statistical data - tabular representation - datasets comparison number 17 [29], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	ZENSUS 2011 - DE - Number of dwellings by household size in Germany (thousands)	CensusHub 2 - 2011 - Number of dwellings by household size in Germany (thousands)
count	80	80
standard deviation	1320.6	1300.7
minimum	4.2	4.0
25%	153.53	154.93
50%	451.08	451.70
75%	1181.08	1171.67
maximum	8127.8	8052.9

Figure 8: Quality control level 3 main statistical data - graphical representation - datasets comparison number 17 [29], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

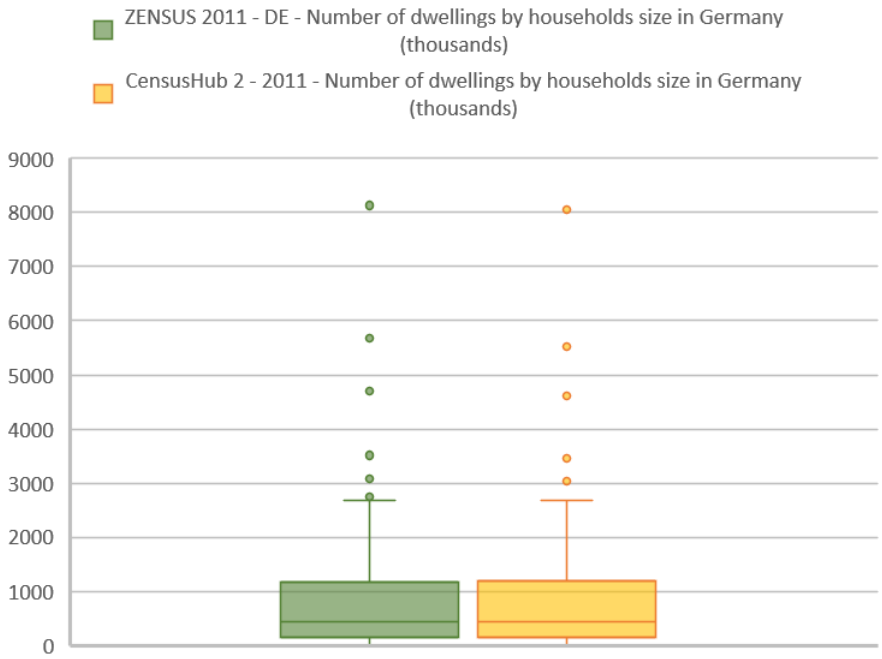


Table 19: Quality control level 3 main statistical data - tabular representation - datasets comparison number 19 [31], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Europe´s buildings under the microscope. A country-by-country review of the energy performance of buildings (BPIE) - Share of owners occupied dwellings [%]	Horizon2020 HotMaps project - Share of owners occupied dwellings [%]
count	15	15
standard deviation	10.4	10.0
minimum	55.0	58.0
25%	66.50	68.00
50%	73.00	73.00
75%	75.50	79.50
maximum	95.0	96.0

Figure 9: Quality control level 3 main statistical data - graphical representation - datasets comparison number 19 [31], [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

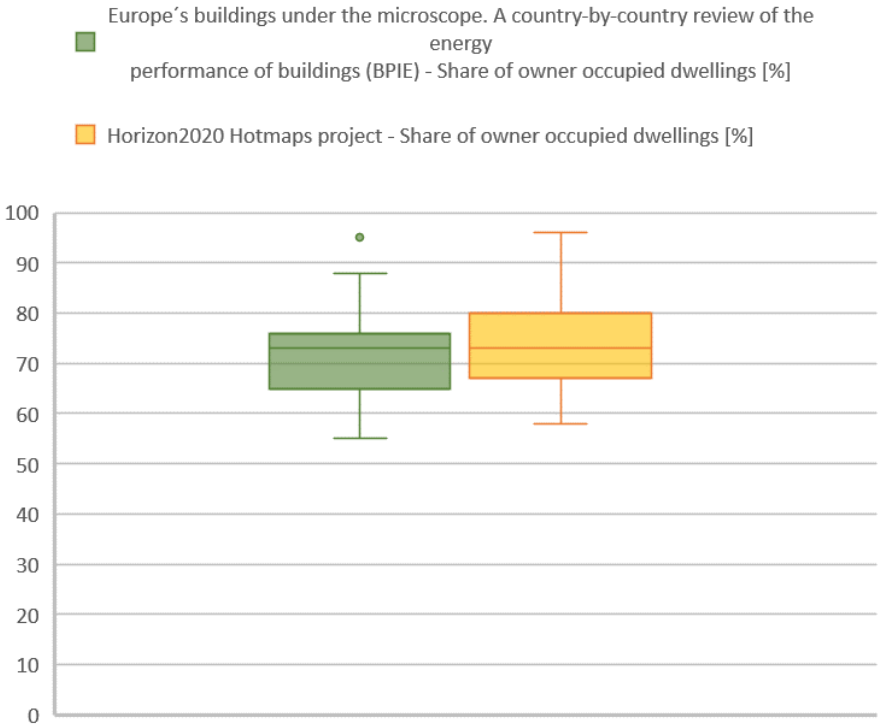


Table 20: Quality control level 3 main statistical data - tabular representation - datasets comparison number 21 [34], [33] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	EDGAR - carbon dioxide emissions in the buildings sector [tCO2]	European Union energy statistical pocketbook - European data portal - carbon dioxide emissions in the buildings sector [tCO2]
count	23	23
standard deviation	32612.2	23953.5
minimum	51.1	50.3
25%	1761.10	1547.15
50%	5223.43	3613.56
75%	26778.86	22252.37
maximum	133181.8	94367.2

Figure 10: Quality control level 3 main statistical data - graphical representation - datasets comparison number 21 [34], [33] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

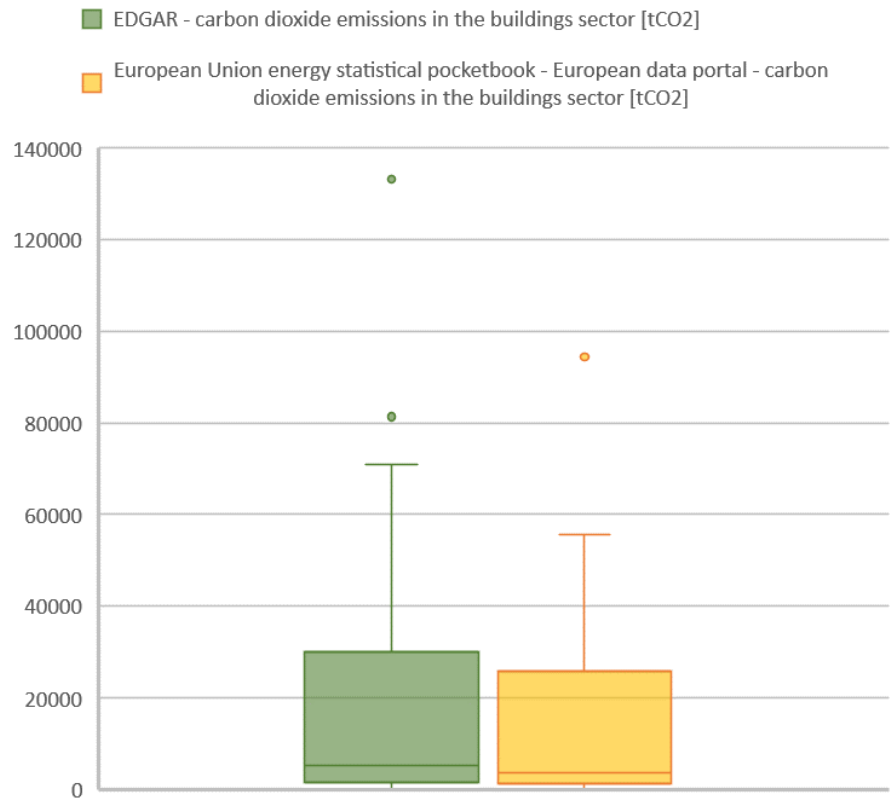


Table 21: Quality control level 3 main statistical data - tabular representation - datasets comparison number 23 [11], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Horizon 2020 HotMaps project - Number of occupied dwellings 2016 [mil]	National Housing Census - Number of occupied dwellings 2011 [mil]
count	26	26
standard deviation	11.8	10.0
minimum	0.2	0.2
25%	1.64	1.53
50%	4.74	3.78
75%	8.59	7.21
maximum	44.0	36.9

Figure 11: Quality control level 3 main statistical data - graphical representation - datasets comparison number 23 [11], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Table 22: Quality control level 3 main statistical data - tabular representation - datasets comparison number 24 [36], [24] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Energy prices in 2019 Household energy prices in the EU increased compared with 2018 - electricity prices in EUR/100kWh in households	2016 - Energy consumption and energy efficiency trends in the EU-28 for the period 2000-2016 - electricity prices in EUR/100kWh in households
count	28	28
standard deviation	5.3	5.5
minimum	9.6	9.5
25%	14.18	13.00
50%	17.90	16.00
75%	21.95	21.00
maximum	29.2	31.0

Figure 12: Quality control level 3 main statistical data - graphical representation - datasets comparison number 24 [36], [24] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

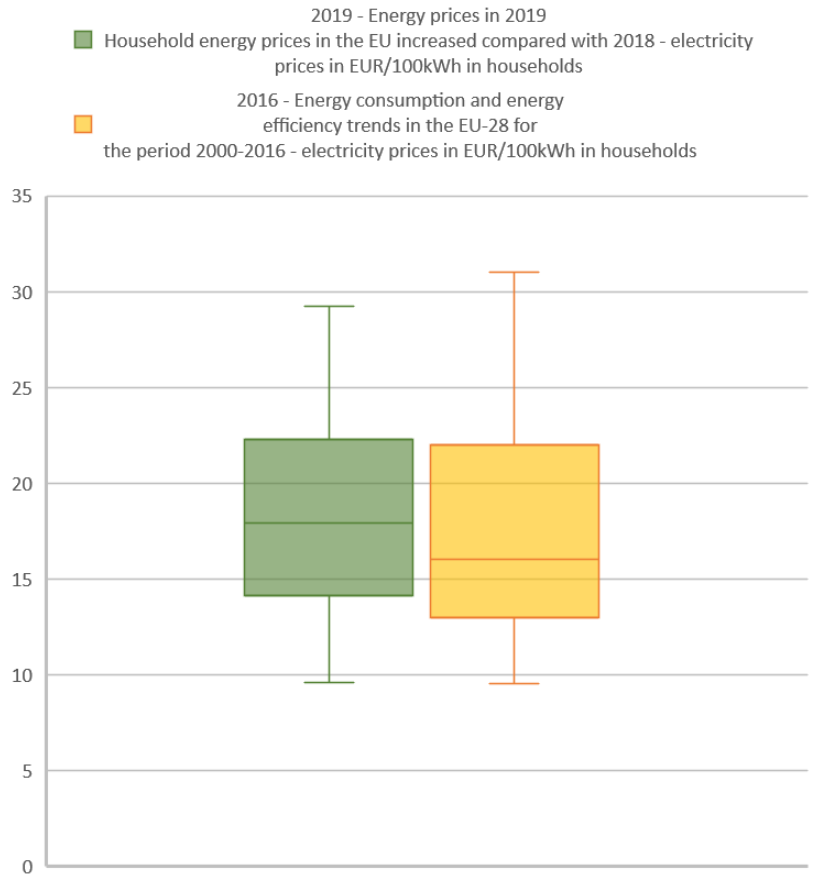


Table 23: Quality control level 3 main statistical data - tabular representation - datasets comparison number 25 [37], [38] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	GDP in PPS by the "World Bank" - 2019	GDP in PPS by EUROSTAT- 2019
count	28	28
standard deviation	41.0	41.3
minimum	52.9	53.0
25%	73.41	73.00
50%	91.77	92.00
75%	117.95	118.25
maximum	260.5	260.00

Figure 13: Quality control level 3 main statistical data - graphical representation - datasets comparison number 25 [37], [38] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

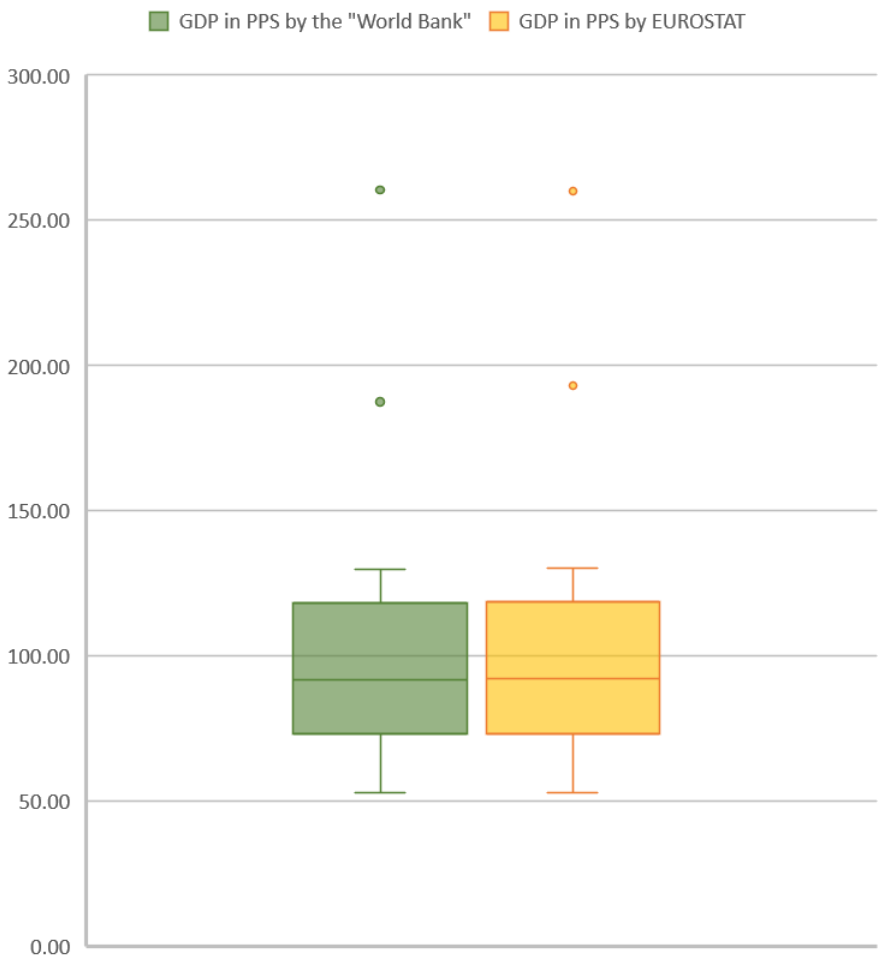


Table 24: Quality control level 3 main statistical data - tabular representation - datasets comparison number 26 [26], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	EUROSTAT - total population 2016	EU Census - total population - 2011
count	226	226
standard deviation	1452894.9	1466511.6
minimum	28 007	28,007
25%	831537.50	828078.50
50%	1346962.00	1342031.50
75%	2142205.75	2149146.75
maximum	11 852 851	12 082 144

Figure 14: Quality control level 3 main statistical data - graphical representation - datasets comparison number 26 26], [17] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

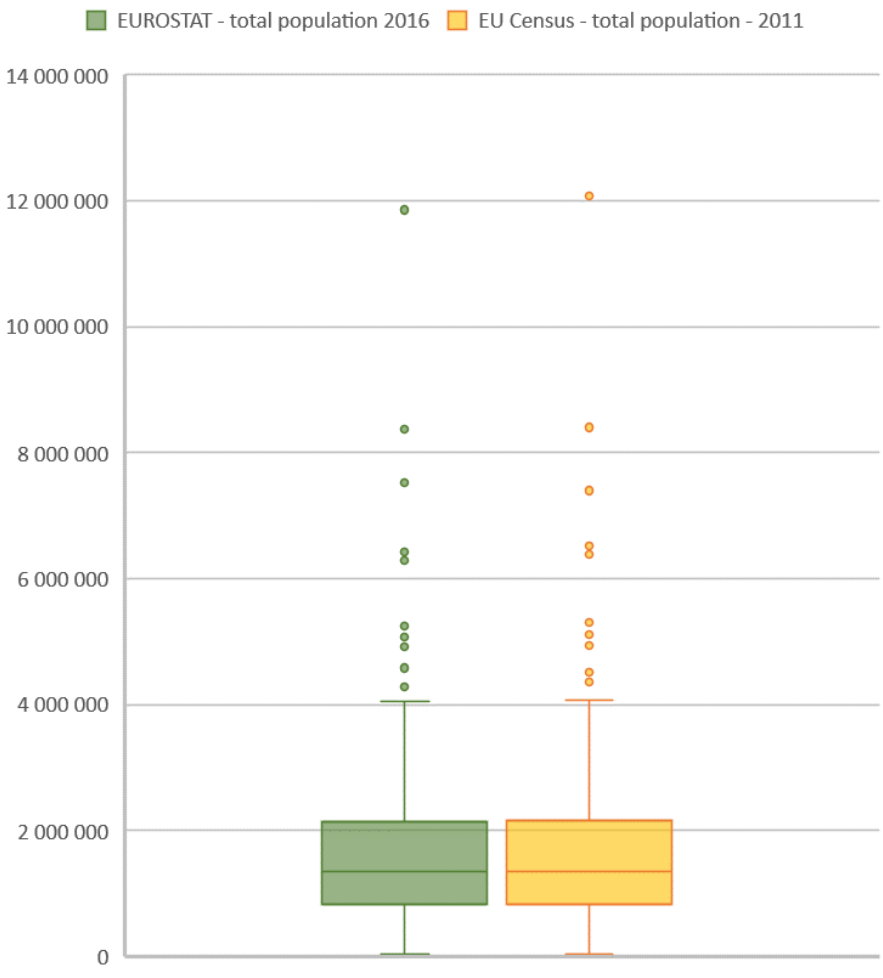
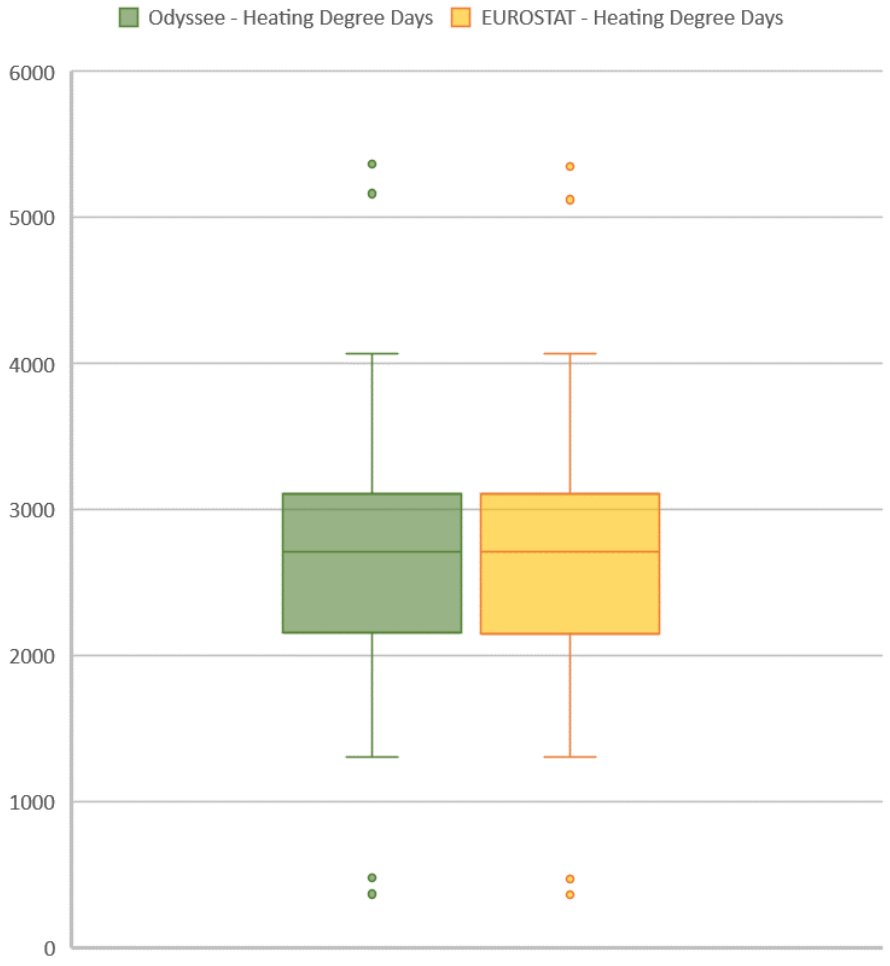


Table 25: Quality control level 3 main statistical data - tabular representation - datasets comparison number 27 [18], [39] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	Odyssee - Heating Degree Days	EUROSTAT - Heating Degree Days
count	28	28
standard deviation	1114.5	1110.1
minimum	366.13	365.57
25%	2174.71	2171.32
50%	2709.32	2707.64
75%	3069.68	3067.26
maximum	5363.53	5349.59

Figure 15 : Quality control level 3 main statistical data - graphical representation - datasets comparison number 27 [18], [39] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].



11. Annex B

This annex entails a number of examples of complete set of data evaluated and compared for quality control of level 3 both in a tabular and in a graphical representation, for whom a difference higher than 30% in one or multiple parameters has been registered.

Table 26: Quality control level 3 main statistical data - tabular representation - datasets compared are [4] and [1] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	IEE Zebra2020 project - Mean U-value Walls 2010-2016 [W/m2K]	Horizon 2020 HotMaps project - Mean U-value Walls 2010-2016 [W/m2K]
count	16	16
standard deviation	0.18	0.18
minimum	0.15	0.18
25%	0.30	0.21
50%	0.32	0.30
75%	0.39	0.38
maximum	0.95	0.78

Figure 16: Quality control level 3 main statistical data - graphical representation - datasets compared are [4] and [1] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

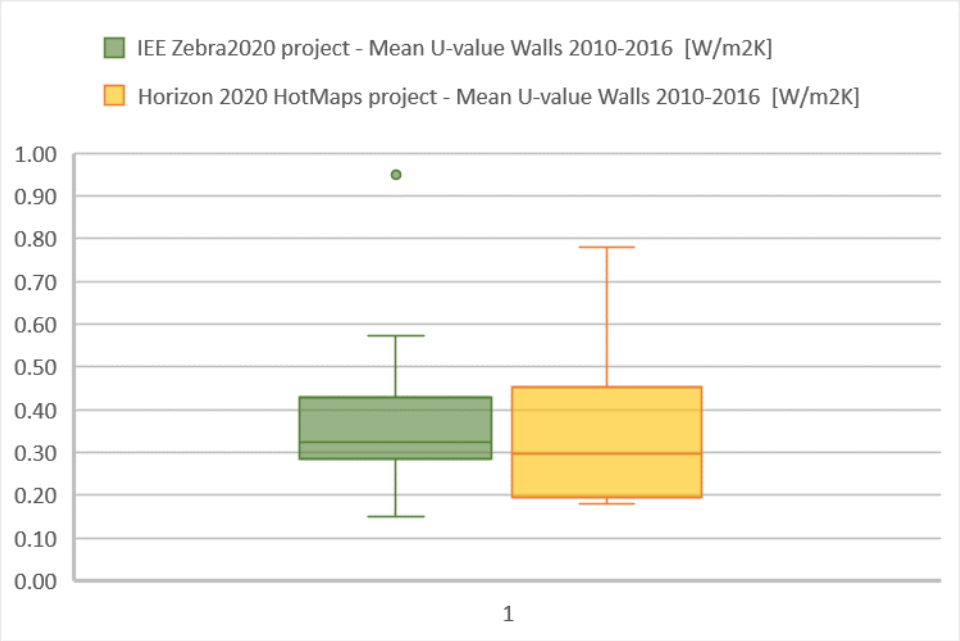


Table 27: Quality control level 3 main statistical data - tabular representation - datasets compared are [1] and [8] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	HotMaps - space heating final energy consumption [Mtoe]	JRC - IDEES - 2015 - space heating final energy consumption [Mtoe]
count	28	28
standard deviation	11.2	8.4
minimum	0.0	0.0
25%	1.14	0.87
50%	3.66	3.13
75%	6.53	5.41
maximum	52.4	35.4

Figure 17: Quality control level 3 main statistical data - graphical representation - datasets compared are [1] and [8] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

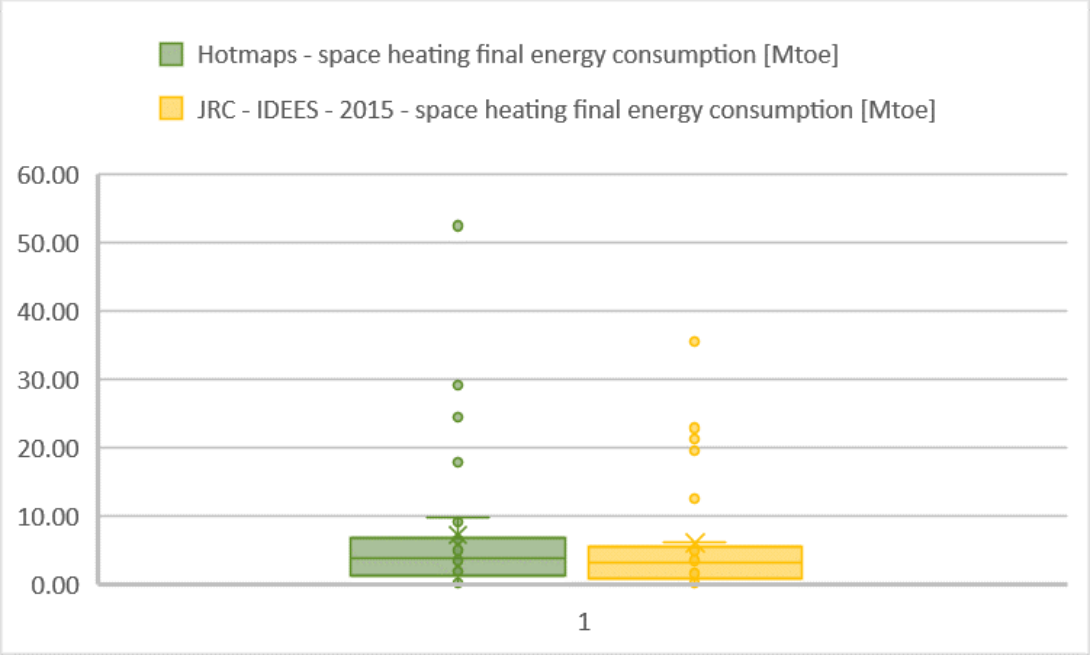


Table 28: Quality control level 3 main statistical data - tabular representation - datasets compared are [8] and [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	JRC IDEES 2015 - Lighting consumption in households (Twh/y)	FP7 Inspire - Lighting consumption in households (Twh/y)
count	27	27
standard deviation	1.6	4.5
minimum	0.0	0.0
25%	0.20	1.00
50%	0.58	2.00
75%	1.14	5.50
maximum	5.9	14.0

Figure 18: Quality control level 3 main statistical data - graphical representation - datasets compared are [8] and [11] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

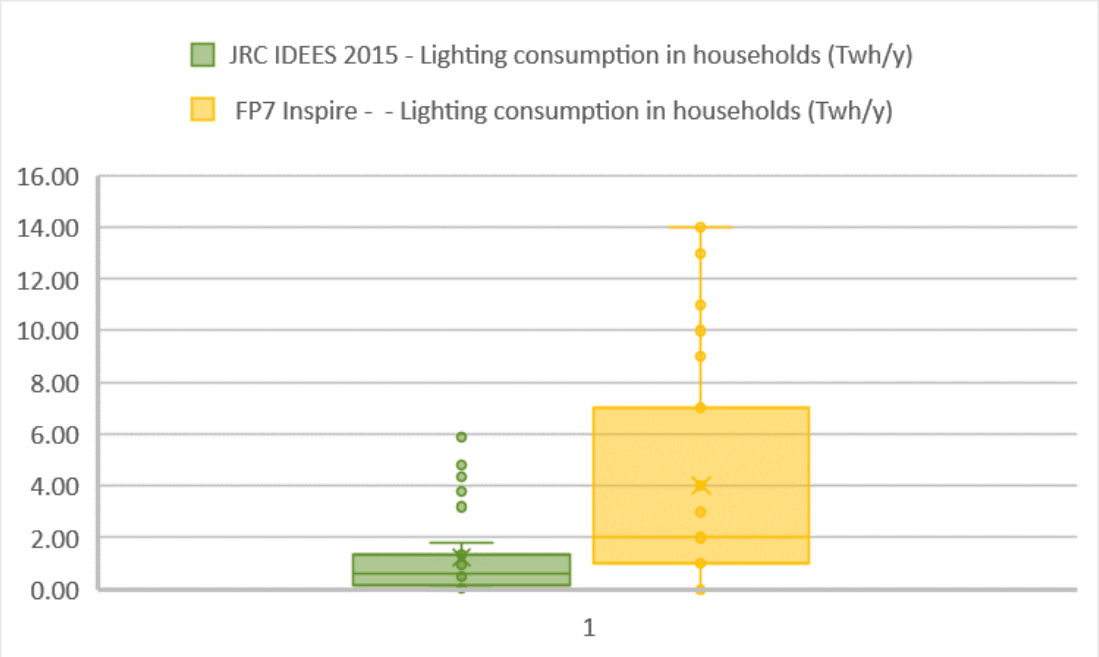
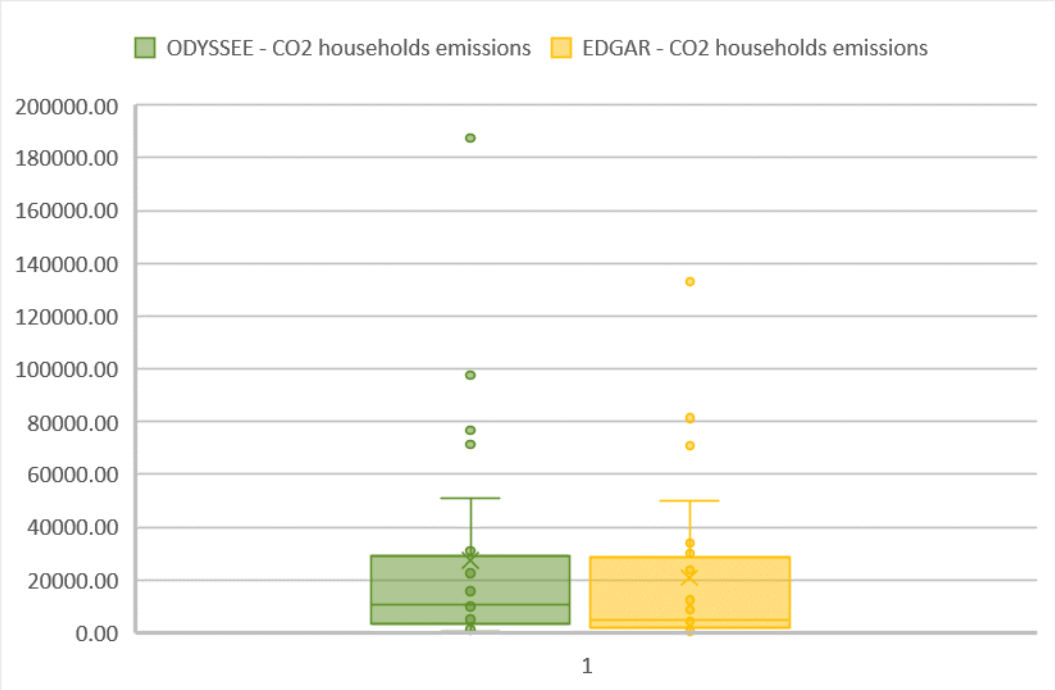


Table 29: Quality control level 3 main statistical data - tabular representation - datasets compared are [6] and [28] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].

Statistical indicators	ODYSSEE - CO2 households emissions [ktCO2]	EDGAR - CO2 households emissions [ktCO2]
count	24	24
standard deviation	42411.3	32111.6
minimum	326.14	51.14
25%	3295.85	1819.62
50%	10380.10	4933.89
75%	25616.85	25181.02
maximum	187477.40	133181.80

Figure 19: Quality control level 3 main statistical data - graphical representation - datasets compared are [6] and [28] according to Table 2. For more information concerning the datasets used please see Deliverable D3.1 of BuiltHub [2].



List of tables

List of figures

Sources

[1] European Commission. BuiltHub Grant Agreement number 957026. 2020.

[2] Zambito A., Pezzutto S., Filippi Oberegger U., Lollini R. and Jankovic I. BuiltHub Deliverable D3.1 - Inventory structure and main feature and datasets. 2021.

[3] Go FAIR. FAIR Principles. 2021. https://www.go-fair.org/fair-principles/

[4] IUCN – International Union for Conservation of Nature. Climate change in Europe. 2020. https://www.iucn.org/regions/europe/our-work/policy/climate-change-europe

[5] European Council. Climate change: what the EU is doing. 2020 https://www.consilium.europa.eu/en/policies/climate-change/

[6] European Commission (EC). A European Green Deal. 2019. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

[7] Build Up Editor. Dynamic EU building stock knowledge hub – Objective and Partners. 2020. https://www.buildup.eu/en/explore/links/builthub-project

[8] European Commission (EC). In focus: Energy efficiency in buildings. 2020. https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-feb-17_en

[9] DataCite metadata scheme. 2020.https://schema.datacite.org/

[10] Schema-org metadata scheme. 2020. https://schema.org/docs/about.htm

[11] Simon Pezzutto, Silvia Croce, Stefano Zambotti, et al. Horizon 2020 HotMaps project: Building stock analysis. 2019. https://gitlab.com/hotmaps/building-stock/-/tree/master/data

[12] Vienna University of Technology - Energy Economics Group (EEG). FP7 CommONEnergy Project: building stock. 2017. https://eeg.tuwien.ac.at/commonenergy/floor-area-non-residential

[13] TABULA Consortium. IEE TABULA project: Typology Approach for Building Stock Energy Assessment. 2017. http://webtool.building-typology.eu/#bm

[14] EPISCOPE Consortium. IEE EPISCOPE project: Focus of building stock monitoring. 2016. https://episcope.eu/welcome/

[15] Zebra2020 Consortium. IEE ZEBRA2020 project: Nearly Zero-Energy Building Strategy 2020. 2015. https://zebra-monitoring.enerdata.net/

[16] ENTRANZE Consortium. IEE ENTRANZE project: Policies to Enforce the TRAnsition to Nearly Zero Energy buildings in the EU27. 2013. https://entranze.enerdata.net/

[17] Statistical Office of the European Union (Eurostat). National Housing Census: European statistical System. 2014. https://ec.europa.eu/CensusHub2/query.do?step=selectHyperCube&qhc=false

[18] ENERDATA. Horizon 2020 ODYSSEE - MURE project: Comprehensive monitoring of efficiency trends and policy evaluation in EU countries, Norway, Serbia and Switzerland. 2020. https://odyssee.enerdata.net/database/

[19] European Commission's Joint Research Centre (JRC). JRC IDEES 2015. 2018. https://data.jrc.ec.europa.eu/dataset/jrc-10110-10001/resource/f590b6f1-60e5-49a6-a972-60bc2b2e34b3

[20] SET-Nav Consortium. SET-Nav - Strategic Energy Roadmap. http://www.set-nav.eu/

[21] Statistical Office of the European Union (Eurostat). Final energy consumption in households by fuel. 2020. https://ec.europa.eu/eurostat/databrowser/view/t2020_rk210/default/table?lang=en

[22] ExcEED Consortium. Horizon2020 ExcEED Project: building stock data. http://www.exceedproject.eu/

[23] iNSPiRe Consortium. FP7 iNSPiRe project: building stock analysis. 2016. https://zenodo.org/record/3256270#.X_wayHqSnIU

[24] European Commission's Joint Research Centre (JRC) - TSEMEKIDI TZEIRANAKI Sofia, BERTOLDI Paolo, LABANCA Nicola et Al. Energy consumption and energy efficiency trends in the EU-27+UK for the period 2000-2016 - FINAL REPORT. 2018. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/energy-consumption-and-energy-efficiency-trends-eu-28-period-2000-2016

[25] Directorate-General for Energy (European Commission) , IPSOS , Navigant. Comprehensive study of building energy renovation activities and the uptake of nearly zero-energy buildings in the EU - FINAL REPORT. 2019. https://op.europa.eu/en/publication-detail/-/publication/97d6a4ca-5847-11ea-8b81-01aa75ed71a1/language-en/format-PDF/source-119528141

[26] Statistical Office of the European Union (Eurostat). Population on 1 January by age, sex and NUTS 2 region. 2020. https://ec.europa.eu/eurostat/databrowser/view/DEMO_R_D2JAN__custom_442767/default/table?lang=en

[27] Statistical Office of the European Union (Eurostat). Final energy consumption in households. 2021. https://ec.europa.eu/eurostat/databrowser/view/t2020_rk200/default/table?lang=en

[28] Statistical Office of the European Union (Eurostat). Disaggregated final energy consumption in households. 2020. https://ec.europa.eu/eurostat/databrowser/view/nrg_d_hhq/default/table?lang=en

[29] Statistische Ämter des Bundes und der Länder 2020. ZENSUS 2011. 2013. https://ergebnisse.zensus2011.de/#

[30] ADEME - Environment and Energy Management Agency. DPE - Diagnostic de Performance Energetique. https://www.observatoire-dpe.fr/index.php/graphique/dpeParEtiquette

[31] BPIE - Europe's buildings under the microscope; Marina Economidou et Al. (BPIE). 2011. https://bpie.eu/wp-content/uploads/2015/10/HR_EU_B_under_microscope_study.pdf

[32] Energy Efficiency Financial Institutions Group (EEFIG). DEEP - De-risking Energy Efficiency Platform. 2021. https://deep.eefig.eu/overview

[33] European Union energy statistical pocketbook. DG Energy. 2019. https://data.europa.eu/data/datasets/eu-energy-statistical-pocketbook?locale=en

[34] Crippa, M., Oreggioni, G., Guizzardi, D., Muntean, M., Schaaf, E., Lo Vullo, E., Solazzo, E., Monforti-Ferrario, F., Olivier, J.G.J., Vignati, E. (Joint Research Centre). EDGAR (Emissions Database for Global Atmospheric Research) CO2 Emissions. 2019. https://github.com/openclimatedata/edgar-co2-emissions

[35] Pezzutto S., Haas F., Exner D., Zambotti S. Dataset of the publication: Europe's Building Stock and Its Energy Demand: A Comparison Between Austria and Italy. 2018. https://link.springer.com/chapter/10.1007%2F978-3-319-75774-2_3

[36] Statistical Office of the European Union (Eurostat). Energy prices in 2019 - Household energy prices in the EU. 2020. https://ec.europa.eu/eurostat/documents/2995521/10826603/8-07052020-AP-EN.pdf/2c418ef5-7307-5217-43a6-4bd063bf7f44

[37] Statistical Office of the European Union (Eurostat). GDP per capita in PPS. 2020. https://ec.europa.eu/eurostat/databrowser/view/tec00114/default/table?lang=en

[38] The world bank. GDP per capita PPP – European Union. 2020. GDP per capita, PPP (current international $) - European Union | Data (worldbank.org)

[39] Statistical Office of the European Union (Eurostat). Cooling and heating degree days. 2020. https://ec.europa.eu/eurostat/web/energy/data/database

[40] European Centre for Medium-Range Weather Forecasts. CORDEX - Regional climate model data on single levels for Europe. 2019. https://cds.climate.copernicus.eu/cdsapp#!/dataset/projections-cordex-single-levels?tab=overview

[41] Climate Monitoring Satellite Application Facility (CM SAF). PVGIS - Photovoltaic Geographical Information System. 2019. https://ec.europa.eu/jrc/en/PVGIS/downloads/CMSAF

[42] Minitab Express support. Interpret the key results for Correlation. 2019. https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/correlation/interpret-the-results/

[43] Statistics by Jim. How to Perform Regression Analysis using Excel. 2021. https://statisticsbyjim.com/regression/regression-analysis-excel/

[44] Building Stock Observatory. https://ec.europa.eu/energy/topics/energy-efficiency/energy-efficient-buildings/eu-bso_en

BuiltHub

builthub.eu