ENER MAPS

THE OPEN DATA TOOL EMPOWERING
YOUR ENERGY TRANSITION.

DELIVERABLE 1.6: QUALITY-CHECK PROCESS RESULTS REPORT

E. Wilczynski, S. Pezzutto (EURAC)

Revised by
D. von Gunten (CREM)

26 February 2021

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

enermaps.eu

Executive Summary

This document is Deliverable 1.6 (D1.6): Quality-Check Process Results report of the Horizon 2020 (H2020) EnerMaps project and was submitted within month twelve (M12) of the project (March 2021). This report is mainly a result of Task 1.2: Quality control process, with some elaboration of results from Task 1.3: Focus group and Task 1.4: Final experts review and feedback.

The quality check (QC) process attempts to increase the reliability of the data by determining the overall quality and transparency of the datasets included in EnerMaps, thus increasing the confidence of its users and contributing to efforts to make energy data more findable, accessible, interoperable, and re-usable (FAIR) (1).

The QC process was a multi-step process which included a variety of activities, comprising of:

- a consultation with external experts with regards to the selected datasets and the QC process,
- an in-depth check and collection of relevant metadata,
- a review of dataset documentation for information on methodology and statistical accuracy,
- an assessment of the completeness of the datasets,
- a consistency analysis to ensure that the data are consistent with related data,
- and a statistical assessment of the data with similar datasets to provide a comparative assessment of the data.

The findings of the QC process that have been described and detailed in this report indicate that the selected datasets are both consistent and offer suitable resources to be verified as being high quality. However, these datasets are lacking in numerous areas that prevent them from being truly “FAIR.” Through integration in the EnerMaps Data Management Tool, these datasets will benefit from a common energy metadata standard and the ability to be compared with and accessed easily from a single repository.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

Table of Contents

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

List of figures

List of tables

List of appendices

List of abbreviations

CSA	Coordination and Support Action
D	Deliverable
EDMT	EnerMaps Data Management Tool
EU	European Union
GA	Grant agreement
GDP	Gross domestic product
H2020	Horizon 2020
M	Month
QC	Quality check / quality control
R&I	Research and innovation
WP	Work package

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

1. INTRODUCTION

The EnerMaps project is a H2020 Coordination and Support Action (CSA) project. The overarching goal of EnerMaps is to improve energy research data management practices. Energy data is currently hard to locate, heavily fragmented, and found in several different repositories. These issues inevitably impair research tasks that require the use of this data, leading to delays in research, higher costs, and a general decrease in efficiency in the energy field. EnerMaps aims to resolve these underlined issues by providing a quality-assessed database of critical energy data that will connect researchers and easily provide data using practices to make the data findable, accessible, interoperable, and re-usable (FAIR) (1).

This internal document is Deliverable 1.6 (D1.6) of the EnerMaps project, which is the Quality-Check (QC) Process Results report that contains a description of the methods and a detailed summary of the results that were obtained from Task 1.2: Quality control process (including insights gained from Task 1.3: Focus group and Task 1.4: Final experts review and feedback). The purpose of the report is to describe the methods used to carry out each step of the QC process and to report on the results that were produced during each step. The report is organised by each step of the QC process. In many cases, results were reported in large tables that have been added to the Appendix at the end of the report. The QC Process Results report was submitted within month 12 (M12) of the project.

This report references several other public and internal deliverables that have been previously produced for Work Package 1 (WP1), including D1.2: Focus group report (2) and D1.3: Experts review (3). These deliverables provided essential feedback which was used in various components of the QC process. In addition, the work carried out for D1.3: Experts review is seen as one of the steps of the QC process (discussed further in Section 3: Collection of experts' feedback. Finally, D1.4: Datasets of the EnerMaps Data Management Tool is also referenced as this report contains the full results of the metadata analysis (discussed further in Section 4: Existence check of relevant metadata). A list of the datasets and their creators can be found in Appendix 1: Title (with hyperlink) and Creator of the final dataset inventory.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

2. BACKGROUND

The purpose of the quality check (QC) process (referred also as the quality control process) is to assess the accuracy and quality of the selected datasets. Before starting the quality check process, the selected datasets were categorized into three levels as defined in the grant agreement (GA). The following table shows the steps that are applied to datasets at each level.

Level 1 includes 20 datasets, primarily those that were identified by experts as part of Subtask 1.1.4: Experts selection of datasets. Level 2 includes 20 datasets that have undergone most of the QC process. As the table below demonstrates, these datasets underwent the entire QC process except for the statistical comparison with related datasets. Level 3 includes 10 datasets that underwent the entirety of the QC process, including the statistical comparisons with similar datasets that were not conducted on Level 2 datasets.

The metadata analysis was largely performed for all 3 levels, with a focus on Level 2 and 3 datasets. All datasets were sent to experts as part of Task 1.4: Final experts review and feedback. Finally, it should be noted that since neither the EnerMaps Data Management Tool (EDMT) nor the Kialo social network (4) have launched prior to the release of this report, the input of feedback from users has not been recorded or assessed.

Table 1. Data levels applied to each quality check process step.

QUALITY CHECK PROCESS STEP	LEVEL 1 DATA	LEVEL 2 DATA	LEVEL 3 DATA
Collection of experts' feedbacks (Task 1.4)	X	X	X
Availability of users feed-back in the Kialo social network (4)	X	X	X
Existence check of relevant metadata		X	X
Methodology check of datasets		X	X
Completeness control of the dataset		X	X
Check of statistical accuracy		X	X
Consistency analysis of the datasets		X	X
Comparison with similar datasets			X

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

3. COLLECTION OF EXPERTS' FEEDBACK

3.1. Methods

Experts were asked to provide feedback based on Tasks 1.1: Inventory, 1.2: Quality control process, and 1.3: Focus group. The full background and results of this step of the QC process can be found in D1.3: Expert review report (3). To summarize, experts were provided with a draft version of D1.4: Datasets of the EnerMaps Data Management Tool and were asked to provide feedback using the following brief questionnaire (Figure 1) as an optional guide:

1. How familiar are you with the selected datasets? (For example, have you used any of the datasets or do you know/trust any of the data providers?)
2. What are your thoughts on the selected datasets with regards to your own work? (For example, do they cover important areas of analysis in your field?)
3. The quality-check process provided a basis for assessing the accuracy, completeness, and consistency of the datasets. Do you feel more confident in the accuracy and quality of the selected datasets since they underwent this process?
4. If possible, please provide a user story from your perspective (i.e. "As a <type of user>, I want <some goal> so that <some reason>.") so that we may better identify ways EnerMaps can provide value to energy research and analysis.

Figure 1. Brief questionnaire provided to experts (3).

3.2. Results

3.2.1. FEEDBACK RELATED TO TASK 1.1: INVENTORY

Feedback was gathered on both the specific datasets and the metadata fields that were collected for them. Experts were familiar with datasets from large dataset providers, including Eurostat (5) and Copernicus (6), and considered these resources as especially important for their research and tasks. In addition, experts stated the importance of spatial data and that they perceive a lack of this kind of data in the energy field.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

In terms of metadata, experts provided information on the fields they expect to see in a typical dataset's metadata, including the title of the dataset, the creator, and the publication date and source.

3.2.2. FEEDBACK RELATED TO TASK 1.2: QUALITY CHECK PROCESS

Expert feedback on Task 1.2: Quality check (QC) process was positive. While experts did not have an impression on what to include in a QC assessment, they concluded that such a process was beneficial in increasing the trustworthiness of the data, especially for less well-known datasets (e.g., the Global Tracking Framework energy efficiency indicator dataset (7) and the S2BOIM biomass supply dataset (8)). Experts felt more assured of the accuracy and overall quality of the datasets due to this process.

Experts indicated that the methodology check was a particularly significant step of the QC process. They felt that this step was a key factor in making the datasets more transparent, as researchers would be able to scrutinize the methodology and determine themselves if the methods were sound. Offering a field dedicated to the brief description of the methodology provides researchers with the ability to quickly understand this aspect of the data.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

4. EXISTENCE CHECK OF RELEVANT METADATA

4.1. Methods

The basis for the metadata assessment was identifying the fields mentioned in the GA for Subtask 1.1.2: Dataset description and metadata assessment. The following table lists the included metadata fields along with a description containing the definitions of each field and/or details on how the fields were collected. As per the GA, all 50 of the datasets underwent the metadata check.

The metadata standards created by DataCite (9) and schema.org (10) were used to establish which metadata fields to consider in the metadata assessment. These fields were later amended following consultations with consortium members and external experts via Task 1.3: Focus group and Task 1.4: Final experts review and feedback, respectively.

Table 2. List of metadata fields and descriptions.

METADATA FIELD	DESCRIPTION
Level	The spatial focus of the data.
Spatial granularity	The spatial resolution of the data (for example, the smallest unit of spatial measure).
Identifier	A unique, persistent code or link that can be used to locate the dataset for collection.
Identifier type	The type of identifier selected for the “Identifier” field.
Creator	The organisation and/or individuals responsible for producing the data.
Object	The title of the dataset’s landing page.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

Publisher	The organisation responsible for housing and disseminating the data.
Publication date	Reports the day, month, and year that the data was published. If the dataset was updated and provided a new date, then the most recent date was used. If only a partial date was given, then the reported value was simply the partial date (for example, just the month and year or just the year)
Publication year	The year that the data was published.
Temporal granularity	The temporal resolution of the data.
Time references	The dates/years which the data refers to. This can be a range of years for longitudinal data or a single year for non-longitudinal data.
URLs	The Uniform Resource Locator (URL) of the dataset or dataset landing page.
Content (keywords)	Keywords that describe the data.
Origin	The source effort of the data (for example, the project that produced the dataset).
Geographic extension	The geographical zone containing the data (differs from “Level” field only for raster data).
Projection system	The projected coordinate system that is used for the dataset (only applies to projected data).
Access conditions	A description of whether the data is open and available for download.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

License	The license detailing the use conditions for the data.
Terms of use	Any brief details on the terms of use for the data.
Availability	Describes where the data is available (if the data is publicly available).
Resource type	The type of object (for example, “dataset”).
Data format	The file format of the dataset.
Size of file	The size of the downloaded dataset (and also the compressed dataset, if applicable).
Other relevant information	Any further information that might be necessary (for example, if a login is required to access the data).

4.2. Results

The full results from the metadata check (including the collected values for each metadata field for each dataset) can be found in D1.4: Datasets of the EnerMaps Data Management Tool, which was a public deliverable of the project submitted within M10 (January 2021) (11). A summary of these results can be found in Table 3 on the following page, which reports the amount of missing data for each metadata field. As is evident in the table, metadata was found and collected for most datasets and for most fields. With the exception of two fields (not including the “other relevant information” field since this field was only populated if further information needed to be noted), all metadata fields were found for 98% of the datasets. The two fields which could not be completed to this extent are the “license” and “terms of use” fields. The information present in these fields is crucial in knowing the accessibility and usability of the data. It is likely that for many datasets, the license or terms of use are explicitly stated in a less-obvious area in the data provider’s repository or website. It is important that data providers make this information clearer and, as a best practice, to include this information in the dataset’s metadata.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

Table 3. Summary results of metadata check.

METADATA FIELD	PERCENTAGE MISSING
Level	0
Spatial Granularity	0
Identifier	2
Identifier Type	2
Creator	0
Object	2
Publisher	0
Publication Date	2
Publication Year	2
Temporal Granularity	2
Time references	2
URLs	2
Content (keywords)	0
Origin	0
Geographical extension	0
Projection system	2
Access conditions	0
License	68
Terms of use	96
Availability	0
Resource type	0
Data format	0
Size of file	0
Other relevant information	76

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

5. METHODOLOGY ANALYSIS OF DATASETS

5.1. Methods

As per the GA, the methodology analysis was only conducted for Level 2 and 3 datasets (however, information was collected for Level 1 datasets, if found). For the methodology analysis, the presence of documentation associated with the dataset was checked for and, if it was available, examined. If documentation providing methodology was not found, other sources were considered to gather insights on methodology (e.g., the “About” page of a data repository or project website). If available, links to the documentation or web pages that describe the methodology, in addition to a brief description of the methodology, were added to the dataset list.

5.2. Results

Links to the dataset methodology were not found for seven of the thirty datasets contained in Levels 2 and 3, however brief descriptions of the methodology were created for all thirty datasets. The full results of the methodology analysis can be found in Appendix 2: Results of the methodology analysis.

The quality of the provided methodology varied from dataset to dataset. Some provide a detailed description of methodology in an easy to find repository or dedicated section in documentation. For example, the building stock dataset produced by the HotMaps project has a section on methodology located in a README file associated with the dataset (12). This is considered a best practice since the README file is provided through a persistent link associated specifically with the dataset and can be included in the metadata of the dataset. It should be noted that other datasets provided less detailed methodologies and, as previously mentioned, several datasets did not include an explanation of the methodology.

Including a description of methodology offers several advantages. First, it offers researchers the ability to determine the quality of the datasets for themselves, thus increasing the reliability of the data. Second, if the methodology is easy-to-find and can be located through a persistent URL, then it can be included in the metadata.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

6. COMPLETENESS CONTROL OF THE DATASET

6.1. Methods

The completeness control, or completeness check, involves assessing the extent of missing data present in each Level 2 and 3 dataset. To check for missing data, each dataset was examined for the presence of blank or null values. For practical reasons, this was performed only on datasets that could be opened in spreadsheet software (i.e., Microsoft Excel). Cells that contained blank/null values were those that either had “empty” cells (i.e., no data), or those with a value that was defined to be a representation for missing data (for example, a colon). The extent to the amount of missing data was reported in the dataset list as a percentage of the whole dataset. Finally, the documentation for the data was also reviewed for any mention on previously existing missing data and how resolving these blanks in data were handled (for example, through a method of extrapolation).

6.2. Results

The full results of the completeness control can be found in Appendix 3: Results of the completeness control. Certain datasets contained significant amount of blank or missing values (in specific cases, over 50% of the cells containing data).

Some insights were discovered when conducting this step of the QC analysis. Different data providers handle the issue of missing data differently. Many providers simply leave the cell blank or include a symbol to denote a missing datapoint. For example, datasets provided by Eurostat will often use a colon (“:”) to denote missing data (one specific example of this is Eurostat’s energy efficiency indicator dataset (13)). This is seen as an optimal practice as it makes parsing and identifying missing data—for example, in Microsoft Excel or with other data analysis software—to be a very simple and straightforward task. Since there is a specific symbol denoting missing data, there is no confusion when identifying missing data for removal, extrapolation, or other action from the researcher. However, other datasets will simply add a value of zero (0) to the cell for a missing numeric datapoint. While this is often logically a missing datapoint (for example, if the data for a specific topic for a select country is unlikely to be equal to zero), this practice can also lead to confusion as to whether the value is missing or is, indeed, equal to zero.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

7. CONSISTENCY ANALYSIS

7.1. Methods

As shown in Table 1, the consistency analysis was applied to Level 2 and 3 datasets. A simple linear regression was used to determine if there was evidence of correlation between the assessed dataset and the related dataset that it was being compared to. The approach used for this analysis is summarized as follows:

1. **Select related dataset**: A related dataset that is linearly correlated with the assessed dataset is selected (for example, electricity consumption and population density of specific areas).

2. **Subset data**: Since many of the assessed datasets contain panel data, the structure of the data is not always suitable for simple analysis. To reduce the data dimensionality, a subset of time or location was extracted from the whole dataset (e.g., one year across multiple locations, or one location over numerous years). For point data, data was amalgamated up to the member state level.

3. **Perform regression**: A simple linear regression is then performed with the assessed data subset set as the independent variable and the related dataset set as the dependent variable. The output is then observed. The p-value indicates if the correlation between variables (in this case, between two compared datasets) is statistically significant. Using a significance level of 5%, the p-value from the regression would indicate whether there is evidence of a statistically significant correlation between the two datasets in the model (the assessed dataset and the compared dataset). When testing the significance of correlation, the null hypothesis is that there is not a significant correlation (or linear relationship) between the two variables, while the alternative hypothesis is that there is significant correlation (or linear relationship) between the two variables. Since we are intentionally testing datasets that should be correlated, the expected outcome of the analysis is that the p-value will be lower than the significance level (i.e., less than 5%), resulting in a rejection of the null hypothesis and accepting the alternative. This would indicate some evidence of correlation between the two data subsets. Since we are only interested in a very basic level of correlation, the entire output from the linear regression analysis (for example the r-squared) was not considered, since the extent or full nature of any evidenced correlation was not of interest.

Since the datasets were analysed using linear regression, the assumption is being made that the relationship between the datasets is linear. Assuming the datasets being compared are related linearly, conclusions can be made for the consistency of the data within itself based on the results of the hypothesis test. Failure to reject the null hypothesis of the hypothesis test (in this case, defined as resulting in a p-value greater than the significance level of 5%) would indicate the datasets are not correlated. If the datasets should logically be linearly correlated, this may indicate inconsistencies within one of the data subsets.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

The datasets were selected to be compared with Level 2 and 3 datasets based on two considerations: they either had to attempt to have a similar focus of measure or they had to be logically correlated. Whenever possible, datasets were compared with one of the 50 datasets selected for the initial dataset list since these had already undergone other components of the QC process. Datasets selected to be used to assess the consistency of the Level 2 and 3 datasets were primarily selected based on the first consideration (they had a similar focus of measure). However, if no datasets provided in the initial dataset list satisfied this consideration, then datasets were selected based on the second consideration (the datasets had to be logically correlated). While selecting datasets based on the first consideration was straightforward, selecting datasets for comparison based on their expected consistency was not as simple. In these cases, datasets were compared with socioeconomic data (for example, regional GDP or population).

7.2. Results

The full results for the consistency analysis can be found in Appendix 4: Results of the consistency analysis. The assessed Level 2 and 3 datasets were all found to be consistent based on the evidence of their correlation with related datasets. This was determined based on the results of the linear regression analysis conducted between each assessed dataset and its related dataset—specifically, the hypothesis test of correlation indicated evidence of correlation between the datasets. Figure 2 shows an example regression output from Excel. In this case, the observed statistic was the p-value for the coefficient (in the example, labelled “Eurostat_RES_share”, which is a data from dataset 5—for full results and dataset indexing please refer to the appendix).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.998689199
R Square	0.997380115
Adjusted R Square	0.997289774
Standard Error	0.008910948
Observations	31

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.876646137	0.876646137	11040.18907	5.40285E-39
Residual	29	0.002302745	7.9405E-05		
Total	30	0.878948881			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.002581363	0.002770259	-0.931812859	0.359126448	-0.008247178	0.003084452	-0.008247178	0.003084452
Eurostat_RES_share	0.010034484	9.55007E-05	105.072304	5.40285E-39	0.009839163	0.010229805	0.009839163	0.010229805

Figure 2. Example Microsoft Excel regression output.

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

8. CHECK OF STATISTICAL ACCURACY

8.1. Methods

As mentioned in the GA, the check of the data's statistical accuracy, or simply referred to as the accuracy check, was only conducted on Level 2 and 3 datasets. This step was performed in the same way as the methodology check. The documentation for the data was examined for any elaboration on if the accuracy of the dataset was assessed. If this information was provided, a summary was added to the dataset list.

8.2. Results

Documentation detailing statistical accuracy was not found for 17 of the 30 assessed datasets. The full results of the accuracy check can be found in Appendix 5: Results of the statistical accuracy check. There was little harmonisation on what methods dataset creators used in assessing the statistical accuracy of their data. However, the purpose of this check was to increase the reliability of the datasets by communicating any extra information provided by the dataset's authors on the data's accuracy to researchers.

Comments on statistical accuracy either provided numerical results, for example via a comparison with similar datasets, or, more commonly, a brief comment on the statistical accuracy. It should be noted that while these comments were added to the results, they do not adequately communicate the statistical accuracy of the data. Therefore, the overall conclusion from this step of the QC process is that dataset creators do not regularly provide assessments of statistical accuracy in the dataset documentation. However, it is possible that many of these datasets are assessed for statistical accuracy and that the results are published in a paper or report.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

9. COMPARISON WITH SIMILAR DATASETS

9.1. Methods

Datasets were analysed using Python version 3.9 (14). The Python script used for analysis can be found in Appendix 6: Python script used to compare similar Level 3 datasets. In short, the script loads comma-separated values (CSV) files of the datasets, removes missing data, and then joins the data by a common geographic level (i.e., country) so that the datasets are comparing only data from locations that they have in common. Summary statistics and a boxplot are then generated for the three datasets for comparison. The CSV files of the datasets contain only a subset of each dataset. For simplification purposes, datasets were referred to by their internal identification number (or the acronym for the source organisation in instances where an external dataset was used).

Extracting a subset of each dataset was performed in order to select a common focus among the assessed dataset and the compared datasets so that the same measured variable is assessed for all datasets. In addition, a common time (i.e., year) was selected, whenever possible. The datasets were only analysed for common locations (i.e., countries), meaning that if one dataset did not have data for a specific country, then the data values for this country would not be analysed for the other datasets. In addition, the data for each data subset needed to have common units of measure. If necessary, conversions were made in Microsoft Excel prior to processing/analysing the data so that a common unit of measure was established.

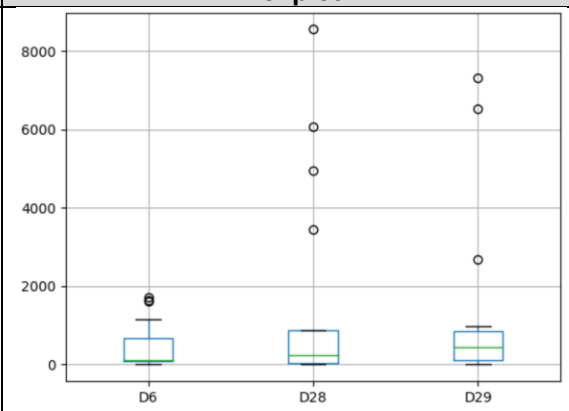
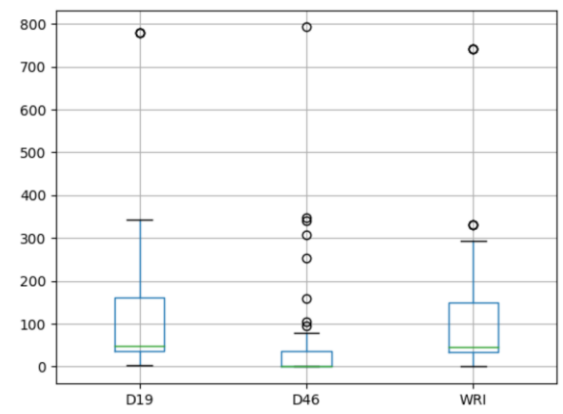
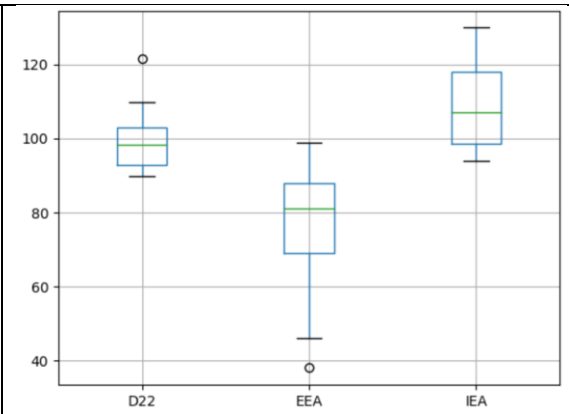
9.2. Results

The full results of the Level 3 dataset comparisons can be found in Table 4 below. Each result has three sections: a statistical summary, boxplot, and brief comment on subset selection. The statistical summary output shows the number of datapoints for the dataset (*count*), the average (*mean*) of the data, the standard deviation (*std*), the minimum (*min*) and maximum (*max*) values, and the first quartile (*25%*), median (*50%*) and third quartile (*75%*) values. The boxplot provides a visual summary of this data.

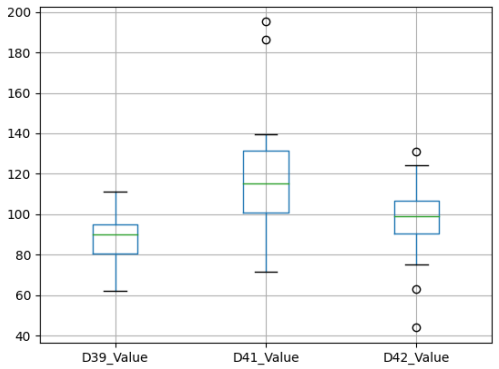
Deviances among the three datasets for each analysis can be attributed to differing collection and/or statistical extrapolation methods. The statistical comparison allows for users to assess which dataset they deem most accurate. For example, a dataset with a mean and variance that is very different than two similar datasets might not be as accurate. A specific example is the first result of the table below, where the means and variance of dataset 6 is quite different than datasets 28 and 29 (for full dataset indexing, please refer to Appendix 1: Title (with hyperlink) and Creator of the final dataset inventory.).

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

Table 4. Results of the Level 3 dataset comparisons with similar datasets.

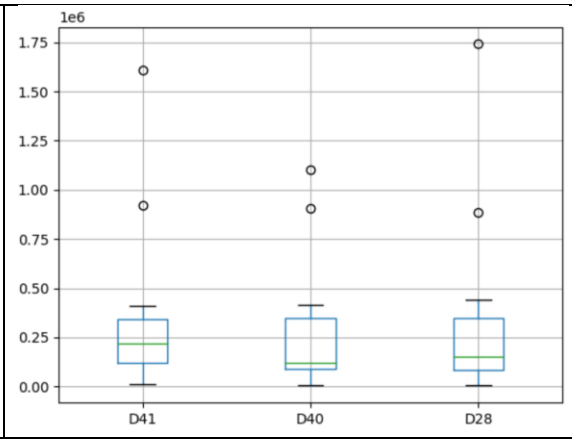
<table>
<tr><th>ID</th><th>Statistical Summary</th><th>Boxplot</th></tr>
<tr><td rowspan="2">6, 28, 29</td><td>
<table>
<tr><td></td><td>D6</td><td>D28</td><td>D29</td></tr>
<tr><td>count</td><td>17</td><td>17</td><td>17</td></tr>
<tr><td>mean</td><td>495.5244</td><td>1521.281</td><td>1272.398</td></tr>
<tr><td>std</td><td>625.6868</td><td>2605.599</td><td>2231.377</td></tr>
<tr><td>min</td><td>7.143061</td><td>0.607884</td><td>4.908031</td></tr>
<tr><td>25%</td><td>84.11507</td><td>45.20255</td><td>123.5</td></tr>
<tr><td>50%</td><td>120.18</td><td>239.422</td><td>446.6765</td></tr>
<tr><td>75%</td><td>683.6744</td><td>871.6883</td><td>849.0229</td></tr>
<tr><td>max</td><td>1710.058</td><td>8560.388</td><td>7333.603</td></tr>
</table>
</td><td> 
</td></tr>
<tr><td colspan="2">Comparison subject: Assessed space cooling consumption in the residential sector in 2015 for 17 countries.</td></tr>
<tr><td rowspan="2">19, 46</td><td>
<table>
<tr><td></td><td>D19</td><td>D46</td><td>WRI</td></tr>
<tr><td>count</td><td>69</td><td>69</td><td>69</td></tr>
<tr><td>mean</td><td>126.8138</td><td>42.79435</td><td>118.8309</td></tr>
<tr><td>std</td><td>174.2987</td><td>119.5212</td><td>165.5725</td></tr>
<tr><td>min</td><td>3.779455</td><td>0.035973</td><td>2.04</td></tr>
<tr><td>25%</td><td>36.23505</td><td>0.081398</td><td>34.86</td></tr>
<tr><td>50%</td><td>48.23717</td><td>0.097783</td><td>45.55</td></tr>
<tr><td>75%</td><td>161.7553</td><td>36.67083</td><td>149.45</td></tr>
<tr><td>max</td><td>780.1188</td><td>792.7932</td><td>740.74</td></tr>
</table>
</td><td> 
</td></tr>
<tr><td colspan="2">Comparison subject: Assessed carbon dioxide emissions in 2014 for 69 countries.</td></tr>
<tr><td rowspan="2">22</td><td>
<table>
<tr><td></td><td>D22</td><td>EEA</td><td>IEA</td></tr>
<tr><td>count</td><td>19</td><td>19</td><td>19</td></tr>
<tr><td>mean</td><td>99.41474</td><td>76.73684</td><td>108.3684</td></tr>
<tr><td>std</td><td>7.833539</td><td>16.97332</td><td>11.61518</td></tr>
<tr><td>min</td><td>89.84</td><td>38</td><td>94</td></tr>
<tr><td>25%</td><td>92.965</td><td>69</td><td>98.5</td></tr>
<tr><td>50%</td><td>98.37</td><td>81</td><td>107</td></tr>
<tr><td>75%</td><td>102.935</td><td>88</td><td>118</td></tr>
<tr><td>max</td><td>121.56</td><td>99</td><td>130</td></tr>
</table>
</td><td> 
</td></tr>
<tr><td colspan="2">Comparison subject: Assessed energy intensity of primary energy consumption in 2015 for 19 countries.</td></tr>
</table>

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

39, 41, 42		

	D39	D41	D42
count	15	15	15
mean	87.26667	120.1813	95.76
std	13.91026	34.6676	22.50536
min	62	71.5	43.9
25%	80.5	100.625	90.5
50%	90	115	99.1
75%	95	131.595	106.55
max	111	195.16	131.1

Comparison subject: Assessed residential floor area for 15 countries.

40		

	D41	D40	D28
count	13	13	13
mean	3.55E+05	2.89E+05	3.39E+05
std	4.41E+05	3.44E+05	4.82E+05
min	1.28E+04	4.85E+03	7.15E+03
25%	1.23E+05	8.87E+04	8.43E+04
50%	2.17E+05	1.22E+05	1.51E+05
75%	3.43E+05	3.50E+05	3.49E+05
max	1.61E+06	1.10E+06	1.74E+06

Comparison subject: Assessed gross floor area of the non-residential sector for 13 countries.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

10. CONCLUSIONS

As discussed in the Section 2: Background, the QC process was performed at three different levels, with each level undergoing more, or less, steps of the QC process. The most comprehensive assessment was the metadata assessment discussed in Existence check of relevant metadata, as this step was applied to all 50 datasets. As concluded in that section, a robust metadata with a full range of necessary metadata fields can be constructed with information provided by the existing metadata or from other resources associated with the dataset (e.g., repository and documentation pages).

Assessing the associated documentation for a dataset did not result in a complete understanding of neither the methodology nor the statistical accuracy for most of the datasets. While this does not necessarily indicate that these datasets are of low quality, the lack of transparency suggests that the user may have to simply trust the quality of the dataset at face value.

Two statistical assessments were carried out for the QC process. The consistency analysis, which was discussed in Section 7: Consistency analysis, determined that the selected datasets of Levels 2 and 3 were consistent when analysed with related data. The statistical comparison with similar datasets, detailed in Section 9: Comparison with similar datasets, demonstrated an additional method to ensure the quality of a dataset. Data is collected and computed in different ways, and so it is important to compare the findings produced in one dataset with others to identify any large discrepancies that may need to be further investigated. Of the analyses performed in this QC step, several of the assessed datasets showed large divergences from similar datasets (specifically, dataset 6 and dataset 46— for full dataset indexing, please refer to Appendix 1: Title (with hyperlink) and Creator of the final dataset inventory.).

The overall conclusion of the QC process is that the selected 50 datasets are consistent and offer enough information and resources to determine that they are high quality. Due to the various collection methods needed to procure the datasets, and their inadequate or non-existent metadata, the selected datasets would benefit greatly from an energy metadata standard and the ability for users to easily access the datasets from a common repository. Therefore, they are suitable for integration into the EnerMaps platform.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

11. APPENDIX

Appendix 1: Title (with hyperlink) and Creator of the final dataset inventory.

ID	**Title** (with Hyperlink)	**Creator(s)**
1	PVGIS: Solar Radiation Data	Climate Monitoring Satellite Application Facility (CM SAF)
2	JRC: Geothermal Power Plant Dataset	Andreas Uihle, Joint Research Centre
3	JRC: Hydro-power plants database	Matteo De Felice, Joint Research Centre, Konstantinos Kavvadias (Joint Research Centre)
4	JRC: Open Power Plants Database	Hidalgo Gonzalez, Ignacio; Kanellopoulos, Konstantinos; De Felice, Matteo; Bocin, Andrei (Joint Research Centre)
5	EEA: Share of gross final consumption of renewable energy sources	European Topic Centre for Air Pollution and Climate Change Mitigation
6	Energy consumption in households	Statistical Office of the European Union (Eurostat)
7	Actual Electricity Generation per Production Type	European Network of Transmission System Operators for Electricity (ENTSO-E)
8	Eurostat: Population	Statistical Office of the European Union (Eurostat)
9	Eurostat: Degree days	Statistical Office of the European Union (Eurostat)
10	COMBI: Annualised net present value of energy efficiency improvement actions	Thema, Johannes; Thomas, Stefan; Teubler, Jens; Chatterjee, Souran; Bouzarovski, Stefan; Mzavanadze, Nora; Suerkemper, Felix; Couder, Johan; Ürge-Vorsatz, Diana; von Below, David
11	SETIS: Private research and innovation (R&I) investment in energy technologies	Pasimeni, Francesco; Fiorini, Alessandro; Georgakaki, Aliki; Marmier, Alain; Jimenez Navarro, Juan Pablo; Asensio Bermejo, Jose Miguel (Joint Research Centre)
12	CORDIS EU research projects under Horizon 2020	European Commission
13	IEA Summary Country RD&D Budgets	International Energy Agency (IEA)
14	Climate Extreme Indices	Mistry, M.N.
15	Copernicus: hourly global climate and weather data	Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N.
16	EMHIRES: Wind power generation	GONZALEZ APARICIO Iratxe; ZUCKER Andreas; CARERI Francesco; MONFORTI Fabio; HULD Thomas; BADGER Jake (Joint Research Centre)
17	EMHIRES: Solar power generation	GONZALEZ-APARICIO Iratxe, HULD Thomas, CARERI Francesco, MONFORTI Fabio, ZUCKER Andreas (Joint Research Centre)
18	Energy Efficiency Indicator	Global Tracking Framework
19	EDGAR CO_2 emissions	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)

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

20	Copernicus: hourly data on pressure levels	Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N.
21	European Digital Elevation Model (EU-DEM)	European Environment Agency
22	Eurostat: Energy efficiency indicator	Statistical Office of the European Union (Eurostat)
23	Projected fresh water use from the European energy sector	Medarac, H., Magagna, D. and Hidalgo González, I. (Joint Research Centre)
24	Photovoltaic power potential	Betak, Juraj; Caltik, Marek; Cebecauer, Tomas; Chrkavy, Daniel; Erdelyi, Branislav; Rosina, Konstantin; Suri, Marcel; Suriova, Nada
25	Wind power density	European Commission and 9 national funding agencies
26	SDH: Large Scale Solar Heating Plants	The European Large-scale Solar Heating Network, Working Group 2E of the European Solar Thermal Technology Platform and the IEE Solar District Heating in Europe project
27	S2BIOM: Biomass supply	S2Biom project consortium
28	HotMaps: Building stock analysis	Simon Pezzutto, Silvia Croce, Stefano Zambotti, EURAC
29	H2020 SET-Nav: Detailed scenario results for energy demand by the INVERT/EE-Lab model	SET-Nav consortium
30	Fuel consumption and technologies used in the heating/cooling sector	Authors of ENER/C2/2014-641 tender
31	INTERREG GRETA	GRETA consortium
32	INTERREG recharge green	recharge green consortium
33	Building Height	European Environment Agency
34	IEA SHC Task 45: Large Solar Heating and Cooling Systems	Sabine Putz, S.O.L.I.D.
35	European Settlement Map	Joint Research Centre
36	TABULA: Building typology data	TABULA project consortium
37	EPISCOPE: Case study scenario analyses	EPISCOPE project consortium
38	ODYSSEE: Gross inland consumption (non-energy uses included)	Enerdata
39	ENTRANZE: Average size of dwelling in residential sector	ENTRANZE project consortium
40	CommONEnergy: Total floor area of the non-residential building sector	Vienna University of Technology Energy Economics Group (EEG)
41	Zebra2020: Share of new dwellings in residential stock	Zebra2020 project consortium
42	National Housing Census: type of living quarter by country	Statistical Office of the European Union (Eurostat)
43	HotMaps: Heat demand density	Andreas Mueller, Mostafa Fallahnejad, TU Wien Institute of Energy Systems and Electrical Drives
44	ECHOES: International survey on energy-related choices and behaviour	Reichl, Johannes; Cohen, Jed; Kollmann, Andrea; Azarova, Valeria; Klöckner, Christian; Royrvik, Jens; Vesely, Stepan; Carrus, Giuseppe; Panno, Angelo; Tiberio, Lorenza; Fritsche, Immo; Masson, Torsten; Chokrai, Parissa; Lettmayer, Gudrun; Schwarzinger, Stephan; Bird, Neil, ECHOES project
45	HotMaps: Heated gross floor area density	Andreas Mueller, TU Wien Institute of Energy Systems and Electrical Drives

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

46	OECD: Greenhouse gas emissions	Organisation for Economic Co-operation and Development (OECD)
47	Electricity prices for household consumers	Statistical Office of the European Union (Eurostat)
48	Expenditure per household on energy	Statistical Office of the European Union (Eurostat)
49	Energy dependence	Statistical Office of the European Union (Eurostat)
50	Regional GDP	Statistical Office of the European Union (Eurostat)

Appendix 2: Results of the methodology analysis.

ID	**Title** (with Hyperlink)	**Methodology** (URL to methodology descriptions)	**Methodology** (brief description)
2	JRC: Geothermal Power Plant Dataset	https://publications.jrc.ec.europa.eu/repository/bitstream/JRC113847/kjna29446enn_jrc113847.pdf	The dataset contains all geothermal power plants that are in operation. Data are collected from various sources and then validated against 3 other datasets. For further information, see the methodology documentation.
3	JRC: Hydro-power plants database	https://github.com/energy-modelling-toolkit/hydro-power-database/blob/master/README.md	The database has been built collecting the information from several other sources and then cross-checking and comparing in case of inconsistencies. For further information, see the methodology documentation.
4	JRC: Open Power Plants Database	https://op.europa.eu/en/publication-detail/-/publication/7930ca56-adbf-11e9-9d01-01aa75ed71a1/language-en	The linkage of available open sources (ENTSO-E, E-PRTR and other power plant databases) enabled the estimation of several performance parameters for a large part of the listed power plants. These are based on analysis of the generation time series provided in ENTSO-E's Transparency Platform, the CO_2 emissions published by the European Environmental Agency's E-PRTR database, as well as the country specific carbon intensity of fuels in each country, published by the UNFCCC. For further information, see the methodology documentation
5	EEA: Share of gross final consumption of renewable energy sources	https://www.eea.europa.eu/publications/renewable-energy-in-europe-approximated	The method used by the European Environment Agency (EEA) features a bottom-up calculation, including hundreds of calculations per Member State, using driving data to estimate GHG emissions. Most of the mathematical methods developed for this purpose can be abstracted from GHG calculations and adapted to estimated energy data. For further information, see the methodology documentation.
6	Energy consumption in households	https://ec.europa.eu/eurostat/documents/3859598/5935825/KS-GQ-13-003-EN.PDF/baa96509-3f4b-4c7a-94dd-feb1a31c7291	Energy consumption in households can be calculated using a variety of methods, principally: business surveys, households surveys, administrative data, or modelling techniques. For further information, see the methodology documentation.
8	Eurostat: Population	https://ec.europa.eu/eurostat/documents/3859598/5916677/KS-RA-11-006-EN.PDF/5bec0655-4a55-	EU Member States use different methods that they consider to be best suited to the administrative practices and traditions of their country. The EU then uses an output-oriented harmonisation of the

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

		466d-9a00-fabe83d54649?version=1.0	collected data. For further information, see the methodology documentation.
9	Eurostat: Degree days	https://ec.europa.eu/eurostat/cache/metadata/en/nrg_chdd_esms.htm#stat_process1554283803348	Heating Degree Days (HDD) index: the severity of the cold in a specific time period taking into consideration outdoor temperature and average room temperature (in other words the need for heating). The calculation of HDD relies on the base temperature, defined as the lowest daily mean air temperature not leading to indoor heating. The value of the base temperature depends in principle on several factors associated with the building and the surrounding environment. By using a general climatological approach, the base temperature is set to a constant value of 15°C in the HDD calculation. Cooling degree days (CDD) index: the severity of the heat in a specific time period taking into consideration outdoor temperature and average room temperature (in other words the need for cooling). The calculation of CDD relies on the base temperature, defined as the highest daily mean air temperature not leading to indoor cooling. The value of the base temperature depends in principle on several factors associated with the building and the surrounding environment. By using a general climatological approach, the base temperature is set to a constant value of 24°C in the CDD calculation. For further information, see the methodology documentation.
10	COMBI: Annualised net present value of energy efficiency improvement actions	https://combi-project.eu/wp-content/uploads/D8.1_tool-guide.pdf	Annualised net value of actions were calculated based on selected actions and countries, annualised investments (annuisation based on selected discount rate), and selected annual impacts. Negative values imply costs of energy efficiency improvement (EEI) actions, positive values gains of EEI actions.
11	SETIS: Private R&I investment in energy technologies	https://setis.ec.europa.eu/sites/default/files/reports/monitoring_r_and_i_in_low-carbon_technologies.pdf	The technology coverage follows the integrated SET Plan structure, showing the links between the Energy Union R&I and Competitiveness priorities, the SET Plan Integrated Roadmap and the 10 SET Plan actions. Trends in patents: The data source is PATSTAT, the Worldwide Patent Statistical Database created and maintained by the European Patent Office (EPO). Private R&I investments: Data are estimated based on financial information from publicly available company statements and patent data from PATSTAT. Public (national) R&I investments: The International Energy Agency (IEA) statistics are the main source of data. For further information, see the methodology documentation.
12	CORDIS EU research projects	/	Information from projects and related organisations funded by the European Union under the Horizon

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

	under Horizon 2020		2020 framework programme for research and innovation are collected and amalgamated monthly.
13	IEA Summary Country RD&D Budgets	https://iea.blob.core.windows.net/assets/3432ae79-1645-4cf1-a415-faa3588e6f29/RDDManual.pdf	Budgets are collected through questionnaires and surveys issued by the International Energy Agency to nations.
16	EMHIRES: Wind power generation	https://publications.jrc.ec.europa.eu/repository/bitstream/JRC103442/jrcreport_20161108_lastversion.pdf	Total Full Load Hours (FLH) were calculated using the ratio between the sums of the energy produced (GWh) and the maximum possible generation (installed capacity(GW)*8760h (GWh)) per country. For further information, see the methodology documentation.
17	EMHIRES: Solar power generation	https://publications.jrc.ec.europa.eu/repository/bitstream/JRC106897/emhirespv_gonzalezaparicioetal2017_newtemplate_corrected_last.pdf	Capacity factors were calculated from the ratio between the sums of the energy produced (GWh) and the maximum possible generation (installed capacity (GW)*8760) per country. For further information, see the methodology documentation.
18	Energy Efficiency Indicator	https://webstore.iea.org/global-tracking-framework-2013#:~:text=The%20Global%20Tracking%20Framework%2C%20a,doubling%20the%20global%20rate%20of	SE4ALL Global Tracking Framework for energy efficiency will: Rely primarily on energy intensity indicators; Use PPP measures for GDP and sectoral value-added; Use primary energy supply for national indicators and final energy consumption for sectoral indicators; Complement those indicators with energy intensity of supply and of the major demand sectors; Provide a decomposition analysis to at least partially strip out confounding effects on energy intensity; Use a five-year moving average for energy intensity trends to smooth out extraneous fluctuations. For further information, see the methodology documentation.
19	EDGAR CO_2 emissions	https://op.europa.eu/en/publication-detail/-/publication/9d09ccd1-e0dd-11e9-9c4e-01aa75ed71a1/language-en	In EDGAR, emissions per country and compound are calculated on an annual basis and sector wise by multiplying the country-specific activity and technology mix data by country-specific emission factors and reduction factors for installed abatement system for each sector. For further information, see the methodology documentation.
22	Eurostat: Energy efficiency indicator	https://ec.europa.eu/eurostat/documents/3859598/5885369/NRG-2004-EN.PDF/b3c4b86f-8e88-4ca6-9188-b95320900b3f	The energy efficiency indicators are derived from energy balances, which are obtained when you convert the natural units in the commodity balances to the chosen energy unit by multiplying by the appropriate conversion equivalent for each of the natural units. For further information, see the methodology documentation.
23	Projected fresh water use from the European energy sector	https://ec.europa.eu/jrc/en/publication/projected-fresh-water-use-european-energy-sector	The projected water used by the energy system is based on the combination of the EU Energy Reference Scenario with water withdrawal and consumption factors for the different processes considered throughout the report. These water factors were gathered from a broad literature review. For further information, see the methodology documentation.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

24	Photovoltaic power potential	https://globalsolaratlas.info/support/methodology	The location-specific information provided by the Atlas involves three main different models: Solar radiation model; Air temperature model; PV power simulation model. Solar radiation and air temperature modeling result in a series of pre-calculated data layers that can be retrieved at (almost) any location on the map. Additional information about a possible PV system type and configuration are used for the PV power simulation, which is calculated on-demand using Solargis internal algorithms and databases. For further information, see the methodology documentation.
28	HotMaps: Building stock analysis	https://gitlab.com/hotmaps/building-stock/-/blob/master/README.md	The data collected in the building stock analysis are used as starting point to calculate the useful energy demand (UED) for space heating (SH), space cooling (SC), and domestic hot water (DHW) for each EU28 MS down to its local level, and to derive scenarios for the future development of the UED.
29	H2020 SET-Nav: Detailed scenario results for energy demand by the INVERT/EE-Lab model	https://www.invert.at/methodology.php	The core of the tool is a nested logit approach, which optimizes objectives of “agents” under imperfect information conditions and by that represents the decisions maker concerning building related decisions. For further information, see the methodology documentation.
38	ODYSSEE: Gross inland consumption (non-energy uses included)	https://www.odyssee-mure.eu/faq/energy-efficiency-methodology/	ODYSSEE data are collected by national partners and checked and harmonized by Enerdata. For further information, see the methodology documentation.
39	ENTRANZE: Average size of dwelling in residential sector	/	Entranze includes data collected from numerous sources--including Odyssee, Building Performance Institute Europe, Tabula, and Eurostat--and and presents them in an online data mapping tool.
40	CommONEnergy: Total floor area of the non-residential building sector	/	The data is built on existing studies and surveys from statistical offices, currently on-going projects on national and European level as well as on the calculations carried out as part of the abovementioned analysis of the commercial building stock at EU level. The scenario development was carried out using the abovementioned data and calculated using a bottom-up approach.
41	Zebra2020: Share of new dwellings in residential stock	/	Data in the Zebra2020 Data Tool are gathered from a variety of national sources, including Statistik Austria for Austrian data, Istat for Italian data, and the Direction générale Statistique et Information écon for French data
42	National Housing Census: type of living quarter by country	/	Data in the National Housing Census are taken from national sources which use a variety of techniques to gather data (i.e. surveys).
46	OECD: Greenhouse gas emissions	/	The data is derived from the National Inventory Submissions 2020 to the United Nations Framework Convention on Climate Change (UNFCCC, CRF

The EnerMaps project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°884161

			tables), and replies to the OECD State of the Environment Questionnaire.
47	Electricity prices for household consumers	https://ec.europa.eu/eurostat/documents/3859598/8048500/KS-GQ-16-106-N.pdf/8d9a943f-b0da-4ac9-bc62-251458d0f498	Electricity prices are collected by a half-yearly questionnaire on electricity prices for households. For further information, see the methodology documentation.
48	Expenditure per household on energy	/	In summary the key concepts captured by the national accounts main aggregates datasets cover the following definitions: GDP - Gross domestic product. GDP at market prices is the final result of the production activity of resident producer units It is defined in three ways: 1. GDP Output approach; 2. GDP Expenditure approach; 3. GDP Income approach.
49	Energy dependence	https://ec.europa.eu/eurostat/documents/3859598/5935825/KS-GQ-13-003-EN.PDF/baa96509-3f4b-4c7a-94dd-feb1a31c7291	Various statistical techniques are used to measure household energy use and energy dependence, including: business surveys, households surveys, use of administrative data, modelling, and in situ measurements. For further information, see the methodology documentation.
50	Regional GDP	https://ec.europa.eu/eurostat/documents/3859598/5937641/KS-GQ-13-001-EN.PDF/7114fba9-1a3f-43df-b028-e97232b6bac5	Regional GDP is valued at market prices by adding the regionalised taxes less subsidies on products and imports, and the Value Added Tax (VAT), to regional gross value added (GVA) at basic prices. For further information, see the methodology documentation.

Appendix 3: Results of the completeness control.

ID	**Title** (with Hyperlink)	**Completeness**
2	JRC: Geothermal Power Plant Dataset	NULL values in 28.2% of cells, though most columns complete (lat/long, year, name, gross cap).
3	JRC: Hydro-power plants database	Blank values in 41.4% of cells, though most columns complete (lat/long, year, id, name, installed capacity)
4	JRC: Open Power Plants Database	Blank values in 53.5% of cells and in 75% of lat/long columns (mainly Italy and Spain).
5	EEA: Share of gross final consumption of renewable energy sources	Dataset has no blank/missing values.
6	Energy consumption in households	Missing values in 37.0% of cells.
8	Eurostat: Population	Dataset has no blank/missing values.
9	Eurostat: Degree days	Dataset has no blank/missing values.
10	COMBI: Annualised net present value of energy efficiency improvement actions	Dataset has no blank/missing values

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

11	SETIS: Private R&I investment in energy technologies	Numerous missing values
12	CORDIS EU research projects under Horizon 2020	Certain fields have missing values
13	IEA Summary Country RD&D Budgets	Blank values in 50.0% of cells, though mainly for data in less recent years.
16	EMHIRES: Wind power generation	Certain datasets from which this dataset was derived contained missing values, which were either completed with gap filling or certain datapoints were removed to create a complete dataset.
17	EMHIRES: Solar power generation	Missing data from derived datasets (i.e. CM-SAF SARAH) were reconstructed to create a complete dataset.
18	Energy Efficiency Indicator	Missing values in 16.3% of cells
19	EDGAR CO_2 emissions	Missing values in <0.8% of cells
22	Eurostat: Energy efficiency indicator	Missing values in 77.3% of cells
23	Projected fresh water use from the European energy sector	Dataset has no blank/missing values
24	Photovoltaic power potential	"Missing records are very rare in the modern satellite and model data inputs. Intelligent gapfilling algorithms are used for gap filling. Historical satellite missions show higher percentage of missing or incorrect data records."
28	HotMaps: Building stock analysis	Incomplete source data was reconstructed by "extrapolating and assembling data from large data tools" and "researching data sourceby-source from single scientific literature fonts as journal papers, conference proceedings and project deliverables"
29	H2020 SET-Nav: Detailed scenario results for energy demand by the INVERT/EE-Lab model	Missing values in 9.7% of cells
38	ODYSSEE: Gross inland consumption (non-energy uses included)	Data complete until 2017 (dataset checked: gross inland consumption -- non-energy uses included)
39	ENTRANZE: Average size of dwelling in residential sector	Data missing for Croatia and Cyprus (dataset checked: average size of dwelling by type)
40	CommONEnergy: Total floor area of the non-residential building sector	Dataset has no blank/missing values
41	Zebra2020: Share of new dwellings in residential stock	Missing data present (difficult to estimate given structure of data)
42	National Housing Census: type of living quarter by country	Missing values in <0.5% of cells (for type of living quarter by country)
46	OECD: Greenhouse gas emissions	Missing values in <0.5% of cells
47	Electricity prices for household consumers	For cost of electrical energy per kWh in Euros: missing data in 4.3% of cells

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

48	Expenditure per household on energy	For current prices: missing data in 9.5% of cells
49	Energy dependence	Missing values in 3.3% of cells
50	Regional GDP	Missing values in 6.6% of cells

Appendix 4: Results of the consistency analysis.

ID	**Title** (with Hyperlink)	**Consistency Analysis**
2	JRC: Geothermal Power Plant Dataset	Compared with *EIA. (2020). Energy production. https://www.eia.gov/international/data/world/total-energy/total-energy-production*. Statistically significant correlation (p-value < 0.05).
3	JRC: Hydro-power plants database	Compared with *EIA. (2020). Energy production. https://www.eia.gov/international/data/world/total-energy/total-energy-production*. Statistically significant correlation (p-value < 0.05).
4	JRC: Open Power Plants Database	Compared with *EIA. (2020). Energy production. https://www.eia.gov/international/data/world/total-energy/total-energy-production*. Statistically significant correlation (p-value < 0.05).
5	EEA: Share of gross final consumption of renewable energy sources	Compared with *Eurostat. (2020). Share of energy from renewable sources. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_ind_ren&lang=en*. Statistically significant correlation (p-value < 0.05).
6	Energy consumption in households	Compared with dataset 38. Statistically significant correlation (p-value < 0.05).
8	Eurostat: Population	Compared with *The World Bank. (2020). Population, total. https://data.worldbank.org/indicator/SP.POP.TOTL*. Statistically significant correlation (p-value < 0.05).
9	Eurostat: Degree days	Compared with *European Environmental Agency. (2020). Global and European temperatures. https://www.eea.europa.eu/data-and-maps/indicators/global-and-european-temperature-10/assessment*. Statistically significant correlation (p-value < 0.05).
10	COMBI: Annualised net present value of energy 9efficiency improvement actions	Compared with dataset 29. Statistically significant correlation (p-value < 0.05).
11	SETIS: Private R&I investment in energy technologies	Compared with dataset 13. Statistically significant correlation (p-value < 0.05).
12	CORDIS EU research projects under Horizon 2020	Compared with *International Monetary Fund. (2019). World Economic Outlook Database: GDP. https://www.imf.org/external/pubs/ft/weo/2019/02/weodata/index.aspx*. Statistically significant correlation (p-value < 0.05).
13	IEA Summary Country RD&D Budgets	Compared with *International Monetary Fund. (2019). World Economic Outlook Database: GDP. https://www.imf.org/external/pubs/ft/weo/2019/02/weodata/index.aspx*. Statistically significant correlation (p-value < 0.05).

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

16	EMHIRES: Wind power generation	Compared with *Global Wind Atlas. (2019). Mean Wind Speed. https://globalwindatlas.info/area/*. Statistically significant correlation (p-value < 0.05).
17	EMHIRES: Solar power generation	Compared with dataset 24. Statistically significant correlation (p-value < 0.05).
18	Energy Efficiency Indicator	Compared with dataset 22. Statistically significant correlation (p-value < 0.05).
19	EDGAR CO_2 emissions	Compared data obtained from *World Resources Institute CAIT Climate Data Explorer. (2019). Country Greenhouse Gas Emissions. http://cait.wri.org*. Statistically significant correlation (p-value < 0.05).
22	Eurostat: Energy efficiency indicator	Compared with dataset 18. Statistically significant correlation (p-value < 0.05).
23	Projected fresh water use from the European energy sector	Compared with dataset 50. Statistically significant correlation (p-value < 0.05).
24	Photovoltaic power potential	Compared with dataset 17. Statistically significant correlation (p-value < 0.05).
28	HotMaps: Building stock analysis	Compared with dataset 40. Statistically significant correlation (p-value < 0.05).
29	H2020 SET-Nav: Detailed scenario results for energy demand by the INVERT/EE-Lab model	Compared with dataset 6. Statistically significant correlation (p-value < 0.05).
38	ODYSSEE: Gross inland consumption (non-energy uses included)	Compared with dataset 6. Statistically significant correlation (p-value < 0.05).
39	ENTRANZE: Average size of dwelling in residential sector	Compared with dataset 28. Statistically significant correlation (p-value < 0.05).
40	CommONEnergy: Total floor area of the non-residential building sector	Compared with dataset 28. Statistically significant correlation (p-value < 0.05).
41	Zebra2020: Share of new dwellings in residential stock	Compared with dataset 28. Statistically significant correlation (p-value < 0.05).
42	National Housing Census: type of living quarter by country	Compared with dataset 28. Statistically significant correlation (p-value < 0.05).
46	OECD: Greenhouse gas emissions	Compared data obtained from *World Resources Institute CAIT Climate Data Explorer. (2019). Country Greenhouse Gas Emissions. http://cait.wri.org*. Statistically significant correlation (p-value < 0.05).

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

47	Electricity prices for household consumers	Compared with dataset 48. Statistically significant correlation (p-value < 0.05).
48	Expenditure per household on energy	Compared with dataset 47. Statistically significant correlation (p-value < 0.05).
49	Energy dependence	Compared data obtained from *CIA World Factbook. (2020). Electricity - Imports. https://www.cia.gov/library/publications/resources/the-world-factbook/fields/255rank.html and EIA. (2020). Primary energy production. https://www.eia.gov/international/data/world/total-energy/total-energy-production*. Statistically significant correlation (p-value < 0.05).
50	Regional GDP	Compared with dataset 23. Statistically significant correlation (p-value < 0.05).

Appendix 5: Results of the statistical accuracy check.

ID	**Title** (with Hyperlink)	**Accuracy**
2	JRC: Geothermal Power Plant Dataset	Comparison of total installed capacity versus similar datasets: 0.3% difference vs Platts; 1.3% difference vs Think Geoenergy; 9.6% difference vs World Resources Institute (WRI).
3	JRC: Hydro-power plants database	Joint Research Centre (JRC) dataset: 1248 unique power plants, World Resources Institute (WRI) dataset: 1918 unique power plants
4	JRC: Open Power Plants Database	Joint Research Centre (JRC) dataset: 809 unique power plants, World Resources Institute (WRI) dataset: 2202 unique power plants
5	EEA: Share of gross final consumption of renewable energy sources	/
6	Energy consumption in households	"The accuracy of the basic data depends on the quality of the national statistical systems and may vary from country to country."
8	Eurostat: Population	Difference between grid data totals (integers) and totals from official statistics ranges from 10 to 1146 people based on country
9	Eurostat: Degree days	/
10	COMBI: Annualised net present value of energy efficiency improvement actions	/
11	SETIS: Private R&I investment in energy technologies	"The SETIS estimations of private R&I are a metric aimed at enabling relative comparisons over time, rather than an accurate account of private investment figures"
12	CORDIS EU research projects under Horizon 2020	/
13	IEA Summary Country RD&D Budgets	/
16	EMHIRES: Wind power generation	As opposed to the IRENA dataset, Renewable.ninja dataset, and the Global Wind Atlas, EMHIRES takes into account wind farm specific power curves for each location which increases its accuracy versus the other datasets.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

17	EMHIRES: Solar power generation	The validation of EMHIRES against power system statistics and time series published by Transmission System Operators shows a very good performance over the countries analysed. EMHIRES is able to capture the variability of solar energy, the seasonality and diurnal cycles and also the peaks and ramps. There is a general slight overestimation of the simulations due to the uncertainties accumulated in the theoretical process of the conversion of radiation into generation.
18	Energy Efficiency Indicator	/
19	EDGAR CO_2 emissions	/
22	Eurostat: Energy efficiency indicator	/
23	Projected fresh water use from the European energy sector	Compared to Eurostat estimates, Joint Research Centre estimates are within the reported range for most countries (discrepancies for Germany, Italy, Netherlands, Poland, and Greece)
24	Photovoltaic power potential	"In most situations the expected uncertainty for annual values will be within ±4% for Global Horizontal Irradiance (GHI) values and ±9% for Direct Normal Irradiance (DNI) values for most of Europe and North America (approx. below 50°N) and Japan."
28	HotMaps: Building stock analysis	/
29	H2020 SET-Nav: Detailed scenario results for energy demand by the INVERT/EE-Lab model	/
38	ODYSSEE: Gross inland consumption (non-energy uses included)	/
39	ENTRANZE: Average size of dwelling in residential sector	/
40	CommONEnergy: Total floor area of the non-residential building sector	/
41	Zebra2020: Share of new dwellings in residential stock	Each country has differing source which could affect comparability between countries.
42	National Housing Census: type of living quarter by country	/
46	OECD: Greenhouse gas emissions	/
47	Electricity prices for household consumers	(For 2007 and onward data) the published prices are based on real invoiced prices that are paid by end-users
48	Expenditure per household on energy	/
49	Energy dependence	"Quantitative assessment of accuracy was not performed by Eurostat"
50	Regional GDP	/

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

Appendix 6: Python script used to compare similar Level 3 datasets.

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt

#***Change the current working directory***
os.chdir('C:/Users/ewilczynski/Documents/DATACOMP/c39')
print("Current working directory: {0}".format(os.getcwd()))
print()

#***Change to name of csv that is being compared***
dataset1 = '6.csv'
dataset2 = '28.csv'
dataset3 = '29.csv'

#Add datasets as dataframes
df1 = pd.read_csv(dataset1)
df2 = pd.read_csv(dataset2)
df3 = pd.read_csv(dataset3)

#Remove null values
df1.dropna(inplace = True)
df2.dropna(inplace = True)
df3.dropna(inplace = True)

#Merge and keep only matching values by country
dfm = df1.merge(df2,on='Country').merge(df3,on='Country')

#List of dtypes to include
include = ['float', 'integer']

#Call describe() function
descm = dfm.describe(include = include)

#Create Boxplot
boxplot = dfm.boxplot(column=['D6', 'D28', 'D29'])

#Output
print(descm)
plt.show()
```

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

12. REFERENCES

1. **European Commission.** Guidelines on FAIR Data Management in Horizon 2020. [Online] 2016. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-datamgt_en.pdf.

2. **Wilczynski, Eric, Pezzutto, Simon, and Balest, Jessica.** *EnerMaps Deliverable 1.2: Focus Group report.* 2020.

3. **Wilczynski, Eric and Pezzutto, Simon.** *EnerMaps Deliverable 1.3: Expert Review Report.* 2020.

4. **Kialo.** Kialo. [Online] 2020. https://www.kialo.com/social-network-26786.

5. **Eurostat.** *Eurostat.* [Online] 2020. https://ec.europa.eu/eurostat.

6. **Copernicus.** *Copernicus.* [Online] 2020. https://www.copernicus.eu/en.

7. **The Global Tracking Framework.** Energy Efficiency Indicator Results. [Online] 2018. https://energydata.info/dataset/world-global-tracking-framework-2017/resource/5ed45e2a-0291-4338-aeda-46da78470aff.

8. **S2Biom.** S2Biom project. [Online] 2021. https://www.s2biom.eu/.

9. **DataCite Metadata Working Group.** DataCite Metadata Schema Documentation for the. [Online] 2019. https://doi.org/10.14454/7xq3-zf69.

10. **schema.org.** Energy. [Online] 2020. https://schema.org/Energy.

11. **Wilczynski, Eric and Pezzutto, Simon.** EnerMaps Deliverable 1.4: Datasets of the EnerMaps Data Management Tool. [Online] 2021. https://enermaps.eu/wp-content/uploads/2021/02/EnerMaps_D1.4_DatasetList_January2021_WEB.xlsx.

12. **S. Pezzutto, S. Zambotti, S. Croce, P. Zambelli, G. Garegnani, C. Scaramuzzino, R. Pascuas, A. Zubaryeva, F. Haas, D. Exner, A. Müller, M. Hartner, T. Fleiter, A. Klingler, M. Kühnbach, P. Manz, S. Marwitz, M. Rehfeldt, J. Steinbach, E. Popovski.** Building Stock EU28. *Hotmaps Project, D2.3 WP2 Report – Open Data Set for the EU28.* [Online] 2018. https://gitlab.com/hotmaps/building-stock/-/blob/master/README.md.

13. **Eurostat.** Energy efficiency indicator. [Online] 2018. https://data.europa.eu/euodp/data/dataset/YIX54AYLew2DOmPqK8dRfQ.

14. **Python.org.** Python. [Online] 2021. https://www.python.org/.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

15. **European Commission.** CORDIS EU research projects under Horizon 2020. [Online] 2018. https://data.europa.eu/euodp/en/data/dataset/cordisH2020projects.

16. **International Monetary Fund.** World Economic Outlook Database: GDP. [Online] 2019. https://www.imf.org/external/pubs/ft/weo/2019/02/weodata/index.aspx.

17. **Global Tracking Framework.** Energy Efficiency Indicator. [Online] 2018. https://energydata.info/dataset/world-global-tracking-framework-2017/resource/5ed45e2a-0291-4338-aeda-46da78470aff.

18. **HotMaps Project.** HotMaps. [Online] 2020. https://www.hotmaps-project.eu/.

19. **Balest, Jessica, et al.** *EnerMaps Deliverable 1.1: User Stories and Prioritization.* 2020.

20. **Ghaljaie, F., Naderifar, M. and Goli, H.** Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. *Strides in Development of Medical Education, 14(3).* [Online] 2017. 10.5812/sdme.67670.

21. *Stakeholder analysis and social network analysis in natural resource management.* **Prell, Christina, Hubacek, Klaus and Reed, Mark.** 2009, Society and Natural Resources, pp. 501-518.

22. *Stakeholder categorisation in participatory integrated assessment processes.* **Hare, Matt and Pahl-Wostl, Claudia.** 2002, Integrated Assessment, Vol. 3, pp. 50–62.

23. **HotMaps Project.** HotMaps Toolbox. [Online] 2020. https://www.hotmaps-project.eu/hotmaps-project/.

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

The Open Data Tool empowering
your energy transition.

WHAT IS ENERMAPS?

EnerMaps Open Data Management Tool aims to improve data management and accessibility in the field of energy research for the renewable industry.

EnerMaps tool accelerates and facilitates the energy transition offering a qualitative and user-friendly digital platform to the energy professionals.

The project is based on the FAIR principle defining that data have to be Findable, Accessible, Interoperable and Reusable.

EnerMaps project coordinates and enriches existing energy databases to promote a trans-disciplinary research and to develop partnerships between researchers and the energy professionals.

Project Coordinator
Jakob Rager, CREM
jakob.rager@crem.ch

Communication Coordinator
Clémence Contant, REVOLVE
clemence@revolve.media

The EnerMaps project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°884161

enermaps.eu