D4.1 Data enhancement methods for building stock

MODERATE

Marketable Open Data Solution for Optimized Building-related Energy Services

Table of Contents

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

List of Figures

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

List of Tables

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

PROJECT DURATION: 1 July 2022 – 31 May 2026

WP: 4 **DELIVERABLE:** 4.1 Data enhancement methods for building stock

LEAD BENEFICIARY: TUWIEN

SUBMISSION DATE: 31.08.2023

DISSEMINATION LEVEL: Public

DUE DATE: draft version M8, final version M15

REVISION HISTORY:

DATE	VERSION	AUTHOR/CONTRIBUTOR[1]	REVISION BY[2]	COMMENTS
24/01/23	Draft	Philipp Mascherbauer - TUWIEN Francesca Conselvan - E-THINK Daniele Antonucci - EURAC Yixiao Ma, Mohsen Sharifi, Lukas Engelen - VITO		
15/02/23			Daniele Antonucci - EURAC Yixiao Ma, Mohsen Sharifi, Lukas Engelen - VITO	EURAC & VITO reviewed each other's work
15/11/23	Final	Philipp Mascherbauer - TUWIEN Francesca Conselvan - E-THINK Daniele Antonucci - EURAC Yixiao Ma, Mohsen Sharifi, Lukas Engelen - VITO		
20/11/23			Daniele Antonucci - EURAC Yixiao Ma, Mohsen Sharifi, Lukas Engelen - VITO	EURAC & VITO reviewed each other's work

Disclaimer: The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Acknowledgements:

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

[1] Name SURNAME, ORGANIZATION

[2] Name SURNAME, ORGANIZATION

Executive Summary

Increasing use of building monitoring systems creates an opportunity for data-driven approaches in the entire built environment. MODERATE provides an open platform where data owners can openly share anonymized data. To enable access to heterogenous data sources on buildings the project standard ontologies are used to describe the metadata of each dataset. Data enhancement methods are used to increase the value of datasets and make them useful for a variety of tools to increase the understanding of the data. In this deliverable, we present the possible methodologies for preparing and enhancing data at on different levels of aggregation. Further we show the standard ontology used to effectively store and label various building data to make the data interoperable between various tools within the MODERATE platform as well as facilitating data exchange and data sharing. We discuss the aggregation and disaggregation of building data to building stock and vice versa.

Introduction

Data collected in work package 3 (WP3)[3] will be further used in WP4 where methods are applied to enhance the data and finally create synthetic datasets, masking all personal information making data GDPR compliant. Provided building data is often incomplete or lacking descriptive metadata information, due to privacy constraints or because the data can not be monitored. One specific example is monitored electricity consumption data. A prime example for this would be smart meter data where it is not possible to monitor behavioral data and enrich the consumption data with this kind of information. Additional data, such as household size or number of inhabitants could be monitored, but is almost never shared due to privacy concerns. To better understand building data, we develop data enhancement methods to label and categorize certain data in order to create realistic synthetic data as a next step. Within this deliverable the methods applied to enhance data collected from WP3 will be described concerning building stock data and measured energy consumption profiles. A special focus is put on ensuring the consistency of the synthetic indicators with each other and with the technical data used for these end-uses. E.g., the indicator of the specific energy consumption for a certain end-use should be consistent with technology data being applied in this type of end-use.

This deliverable is structure as follow:

- Chapter 1 presents the standard ontology adopted.
- Chapter 2 provides an overview of existing literature and methods relevant for preparing information on load profiles on building level for labeling.
- Chapter 3 describes the methodologies for aggregating building data on different levels.
- Chapter 4 describes the link of this work package to other work packages in the Moderate project.

[3] WP3 deals with data collection from various sources such as public databases, literature, satellite images etc.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

1 Standard Ontology

Data take on a value when they are correctly described (metadata). In the era of big data, the main problem is interoperability between different datasets. In fact, the disaggregation of information or the incorrect description of it leads to its difficult use, reducing the generation of knowledge from the data. In the building domain, various actions have led to the realization of ontologies and semantics that help to better describe both the dataset and the system from which the data was acquired (e.g. building monitoring system, Building Management System, Technical building management, etc.).

Such ontologies allow for an exchange of information at both the machine and human level, enabling easy interrogation of the dataset from which the data is to be obtained. In MODERATE, this approach is indispensable, both for internal data management and to facilitate a possible relationship with other datasets structured using the same ontologies.

The approach with the 'brick schema' ontology is shown below, but will not be the only one as there are several initiatives in place such as FIWARE, NSGI-LD, Haystack, etc. These will be gradually evaluated within the project, providing case studies of their application as well as a detailed description of their use.

1.1 Brick schema[4]

Nowadays, buildings are increasingly becoming incubators of data and information. The integration of intelligence systems, sensor and networking (i.e. Internet of Things - IoT) in buildings is becoming more and more common. Even though the amount of data generated by smart buildings is growing exponentially, there is still no clear industry-wide standard for using, sharing and exchanging information in a unified way.

Indeed, some of the day-to-day actions applied in construction, such as energy audits, optimization of controls or detection of faults in building systems are often slowed down by the lack of standardization of metadata. This makes processes prohibitively time-consuming and burdensome (from a labor point of view) and not reusable in other applications. This problem is generally related to the lack of semantic interoperability.

The latter is defined by Pritoni et all.[5] as “the capability of two or more networks, systems, devices, applications, or components to work together, and to exchange and readily use information securely, effectively, and with little or no inconvenience to the user”. On the technical side, the interoperability between devices is achieved using the same communication protocol, on the contrary the semantic layer is not defined or unambiguously defined, not allowing the development of applications that can be used in different buildings.

All this shows that it is extremely important to define a univocal semantic layer, which, based on a standard model, allows interoperability between different services and platforms.

The semantic model is a metadata schema that describes precisely and unambiguously the different elements that characterize the building and its systems. The peculiarity is that it identifies different entities by means of a glossary or dictionary and links them to each other using relationships. Ontologies establish the domain’s concepts and relationships, classes, and attributes.

[4] https://brickschema.org/#home

[5] Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis – energies- April 2021 DOI: https://www.mdpi.com/1996-1073/14/7/2024

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

As written in [6] "The World Wide Web Consortium (W3C) established standards that created the Semantic Web, an extension of the World Wide Web aimed to make internet data machine- readable. Ontologies that comply with W3C standards use triples in the form of subject–predicate–object to encode knowledge, following the Resource Description Frame- work (RDF) data model. When multiple triples are put together, they form a directed multigraph. The W3C also provides a set of fundamental languages that can be leveraged to define ontologies using classes and properties (i.e., Resource Description Framework Schema or RDFS), description logics (i.e., Web Ontology Language or OWL) and constraints (i.e., OWL and Shapes Constraint Language or SHACL). Ontologies and Semantic Web technologies have experienced some adoption for internet services, providing interoperability of digitized data, for example, between search engines, web crawlers, and other web-based software"

On the building domain the ontologies developed and under development are summarized in the following schema (see *Table* 1).

Table 1: Metadata schema as resulting of review process of existing applications

Phase of the Building Life Cycle	Group	Schemas
Design and energy modelling	Software	• Industry Foundation Classes (IFC) • Green Building XML (gbXML) • ifcOWL • Tubes • SimModel Ontology • Energy-ADE
Operations	Sensor network, Internet of things (IoT) and smart homes	• Semantic Sensor Network/Sensor, Observation, Sample and Actuator (SSN/SOSA) • Web Thing Model • OneM2M[7] BaseOntology's • One Data Model (oneDM) • Smart Energy Aware Systems • ThinkHome • Building Ontology for Ambient Intelligence • DogOnt • Ontology of Smart Building • Smart Application REFerence (SAREF)
Operations	Commercial building, automation and monitoring	• Project Haystack • BASont • Haystack Tagging Ontology (HTO) • Brick Schema • Google Digital Building Ontology • Semantic BMS Ontology

[6] Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis – energies- April 2021 DOI: https://www.mdpi.com/1996-1073/14/7/2024

[7] https://www.onem2m.org/

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Phase of the Building Life Cycle	Group	Schemas
		- CTRLont - Green Button - RealEstateCore (REC) - Building Topology Ontology (BOT) - Building Automation and Control Systems (BACs) - Knowledge Model for City (KM4City) - EM-KPI Ontology
Operations	Grid-interactive efficient building (GEB) applications	- Facilty Smart Grid Information Model - RESPOND
Operations	Occupants and behaviour	- DNAs Framework (obXML) - Occupancy Profile (OP) Ontology - Onto-SB: Human Profile Ontology for Energy Efficiency in Smart Building - OnCom
Operations	Asset management and audits	- Building Energy Data Exchange Specifications (BEDES) - Virtual Building Information Systems (VBIS) - Ontology of Property Management (OPM)

The core concept of an ontology applied on a building domain is defined in the following table.

Table 2: Main core concepts of building ontology

Category	Concept	Proprieties	Relationship to/from
Zones and Spaces	Space	Function Floor Area	Composed of spaces Adjacent spaces
	Zone	Floor area	Overlaps one or more spaces Overlaps other zones
	Building floor	Orientation	Composed of spaces
Envelope	Envelope element	Type of envelope element(wall, roof, floor, window) Envelop characteristics (e.g., thermal resistance, storage, solar seat gain coefficient)	Part of space
Building System and Equipment	System	Type of system	Composed of components
	Equipment	Type of equipment Rated power draw Rated efficiency	Serves zone Located in space Metered by meter

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Category	Concept	Proprieties	Relationship to/from
		Remaining lifespan	Connected to equipment
	HVAC equipment	Rated capacity	
	Lighting equipment	Rated(max.) luminous flux Minimum relative light output Rated (max.) power Correlated color temperature Spectral power distribution Rated Input voltage Rated (max.) input current	Serves zone/space Located in space Metered by (internal/external) Meter Connected to electrical Junction box or other equipment
	Other end use	Type of end-use	
	Component	Type of component	Part of system Located in space Connected to component
Control Devices	Control device		Has points
	Control point	Input/Output Type Physical/Virtual type Type of virtual point (setpoints, command, alarm) Unit of measure Control interval	Linked to sensor/actuator Linked to time series data
	Control strategy	Schedule Event	Has inputs Has outputs Linked to sensor Linked to actuator Linked to time series data
Sensor/Actuator	Sensor	Type of sensor Unit of measure Measurement Interval Reporting Interval	Senses/Measures point Senses/Measures equipment Aggregates measurements
	Actuator	Unit of measure Actuation interval	Actuates point Actuates equipment Integrates/Prioritizes actuations

Brick is a metadata scheme that takes from the Haystack project the use of tags to preserve the flexibility and ease of use of annotating metadata. Brick unlike Haystack schema places restrictions to

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

prohibit arbitrary tag combinations and relationships. For example, the unit for temperature to be chosen can be only Fahrenheit or Celsius or give an error if sensor and set point occurring together in a tag combination for a data point.

Brick introduces the concept of tag set that group together relevant tags to represent an entity. They are:

1. *Points are physical or virtual entities that generate time-series data. Physical points include actual sensors and setpoints in a building, whereas virtual points encompass synthetic data streams that are the result of some process which may operate on other timeseries data, e.g. average floor temperature sensor.*
2. *Equipment: Physical devices designed for specific tasks controlled by points belonging to it. E.g., light, fan, Air Handling Unit (AHU).*
3. *Location: Areas in buildings with various granularities. E.g. room, floor.*
4. *Resource: Physical resource or materials that are controlled by equipment and measured by points. An AHU controls resources such as water and air, to provide conditioned air to its terminal units.*
5. *Together with these entities, Brick defines a minimal set of relationships that capture the connection between them. A Brick building model can be visualize using the Resource Description Framework (RDF) which represents graph-based knowledge as tuples of (subject, predicate, object) termed triples.*
6. *Unlike the other languages for the building metadata scheme, Brick is distinguished for:*
7. *Completeness: The current version of Brick covers the 98% of the vocabularies found in six buildings in different countries.*
8. *Vocabulary Extensibility: The structure of Tags/TagSets allow easy extensions of TagSets for newly discovered domains and devices while allowing inferences of the unknown TagSets with Tags.*
9. *Usability: Brick represents an entity as a whole instead of annotating it. It promotes consistent usages by different actors. Furthermore, its hierarchical TagSets structure allows user queries more generally applicable across different systems.*
10. *Expressiveness: Brick standardizes canonical and usable relation-ships, which can be easily extended with further specifications.*
11. *Schema Interoperability: Using RDF enables straightforward integration of Brick with other ontologies targeting different domains or aspects.*

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

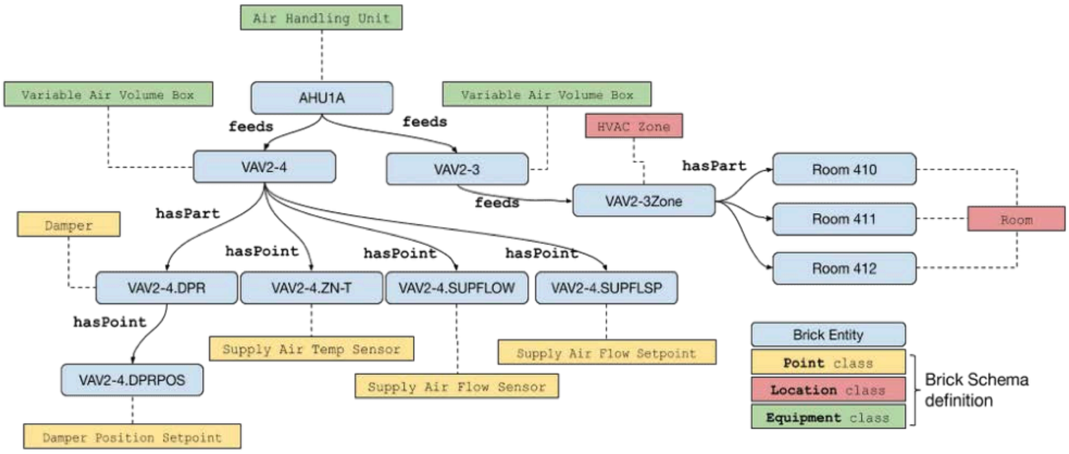


Figure 1: Example of Brick Schema

The relationship supported by Brick and updated by Brick+ are shown in the following table:

Table 3: Relationship and definition for brick and brick plus schema

Relationship	Definition
hasLocation	Subject is physically located in the object entity
feeds	Subject conveys some media to the object entity in the context of some sequential process
hasPoint	Subject has a monitoring, sensing or control point given by the object entity
hasPart	Subject is composed – logically or physically – in part by the object entity
Measures	Subject measures a quantity or substance given by the object entity
Regulates	Subject informs or performs the regulation of the substance given by the object entity
hasOutputSubstance	Subject produces or exports the object entity as a product of its internal process
hasInputSubstance	Subject receives the object entity to conduct its internal process

1.2 SAREF

The Smart Applications REFerence (SAREF)[8] is one of the well-established Ontologies, which is intended to cover the various actors in the Internet of Things (IoT).

[8] https://saref.etsi.org

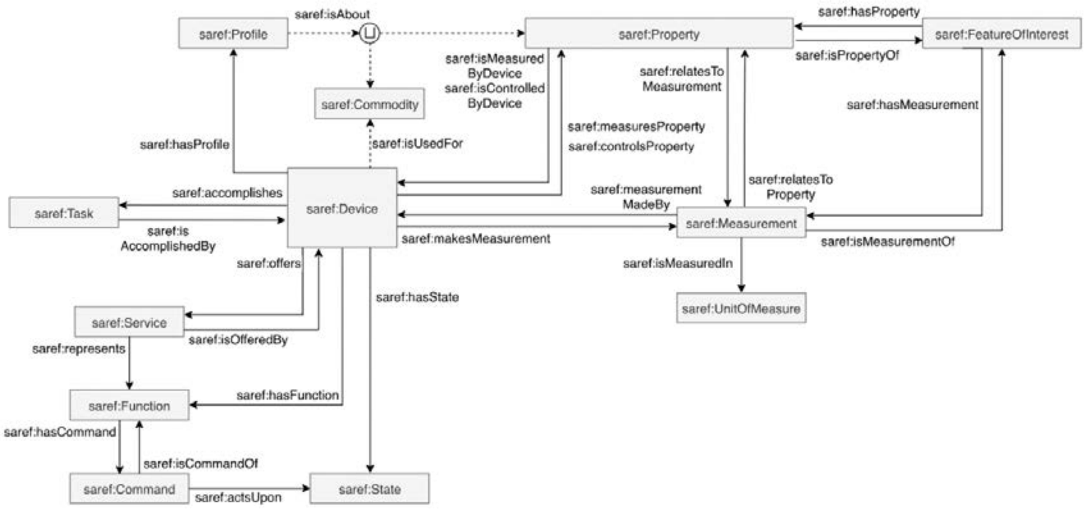


Figure 2: shows an overview of the SAREF. The main classes are contained in the box. The connection between each class is the relationship[9]

SAREF4BLDG is an extension of SAREF, which is created based on the Industry Foundation Classes standard for building information [10]. The goal of SAREF4BLDG is intended to improve the interoperability among different phases of the building life cycle. The overview of SAREF4BLDG is depicted in Figure 3. As can be observed, saref:device is reused from SAREF. The class geo:SpatialThing is from the geo ontology, which proposed the conceptualization for location.

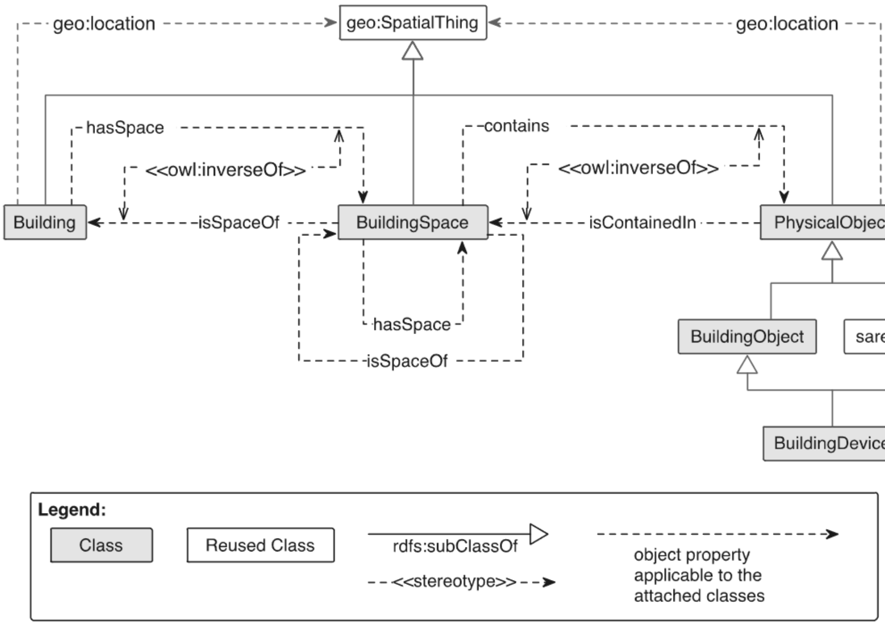


Figure 3: General overview of the top levels of the SAREF4BLDG

[9] https://www.etsi.org/deliver/etsi_ts/103200_103299/103264/03.01.01_60/ts_103264v030101p.pdf

[10] https://www.etsi.org/deliver/etsi_ts/103400_103499/10341003/01.01.02_60/ts_10341003v010102p.pdf

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

2 Load profiles on building level

In this chapter we describe the process of labeling unlabeled load profiles in order to effectively process them later and ultimately create synthetic load profiles.

2.1 Data preparation

The dataset furnished by the energy provider includes clean and high-quality data. All entries come with a timestamp and corresponding energy load details. To facilitate our work, we outline the attributes of the timestamp by extracting the subsequent features: year, month, day, day of the week, week number, classification of holiday or workday, and the current season. Due to changes in winter and summer time we have duplicate and missing values in each profile. The hour corresponding to the missing hour is treated as missing to ensure that records that are not consecutive are not treated as such. Regarding duplicated records, the most recent one is removed. Lastly, we prepared the data set for feature extraction by normalizing the data and making them of the same magnitude for more accurate calculations. We dropped days without a full day record and considered that some days and months are overrepresented since the dataset spans over a period of one and half years.

Electricity consumption contains seasonal patterns, and this affects the calculation of some features, like the measures of central tendency. So, we normalized the data by calculating the mean consumption of each day and averaging it over the entire year. It is important to keep in mind if energy consumption shows a marked long-term trend, that may have an unwanted effect. As well, mean hourly consumption values for each day of the year may exhibit atypical patterns, as they may average out workdays with holidays or days with very different weather conditions and this can artificially reduce the dispersion indices.

2.2 Feature engineering

In the subsequent section, we will delve into the fundamental techniques of feature engineering used to better understand the relationships and the patterns of the energy load profiles. Feature engineering involves transforming raw data into a set of defined features that represent the characteristics of interest. In the context of energy load profiles, this means the utilization of diverse techniques that capture essential aspects of consumption behavior. Six key categories of feature engineering techniques take center stage: central tendency, dispersion, changing in hourly load, zero consumption, peaks and workday and holiday patterns.

Central tendency aids in identifying shifts in consumption patterns over time, facilitating anomaly detection and load forecasting. **Dispersion** highlights the spread or variability in energy load profiles and assesses the stability of energy demand. **Changing in hourly load** underlines fluctuations between consecutive hours and it is a significant value for optimizing generation scheduling.

Zero consumption identifies inactive periods or potential equipment malfunction. **Peak features** identify heightened demand, often revealing peak usage times, and encompass peak load magnitude, frequency, and duration. **Workday vs holiday patterns** involve creating features that distinguish consumption patterns on regular workdays, weekends, and holidays, enabling the model to capture usage fluctuations.

By curating, constructing, and selecting pertinent features, we can uncover hidden patterns, capture inherent dependencies, and facilitate the development of robust predictive models.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

2.2.1 Measure of central tendency

In the context of energy load profiles, the mean and the median are both measures used to understand the central tendency of a set of energy consumption of load data. They provide insights into the typical or average level of energy consumption over a given period.

Mean and median consumption on an hourly and daily basis

The mean consumption on an hourly basis is calculated by taking the mean of all respective hours of each day of the whole profiles (Figure 4). We did not calculate these statistical features on normalized data to preserve the information on the overall consumption. Nevertheless, a comparison with normalized data shows a similar trend for the overall dataset.

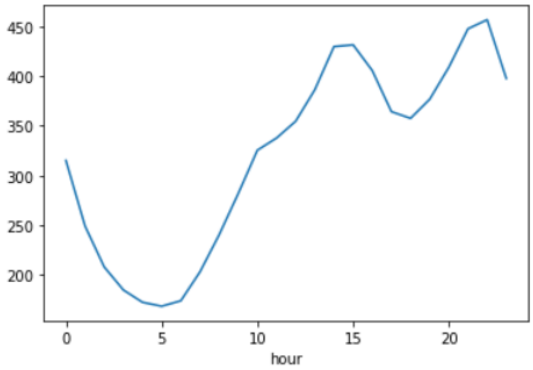


Figure 4: Mean consumption of all consumers within a dataset for every day. The trend shows an increase in consumption between 12 and 3 o'clock and in the evening after 8 pm.

The mean and the median at hourly consumption shows the daily trend, whereas the daily consumption provides information on the consumption pattern throughout the week or the month (Figure 5 and Figure 6).

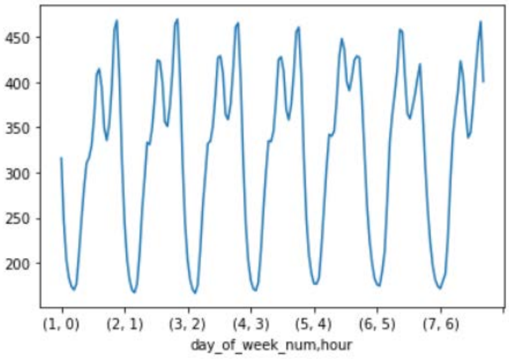


Figure 5: Mean consumption over an average week. Mean consumption of all consumers within the dataset for every day.

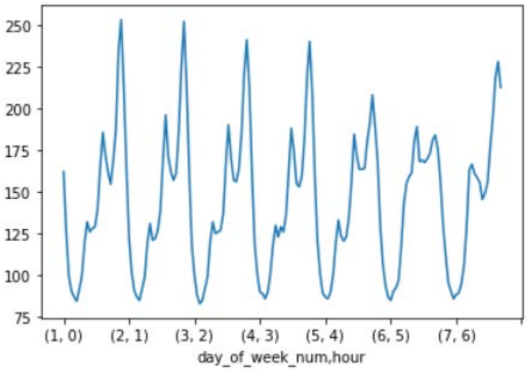


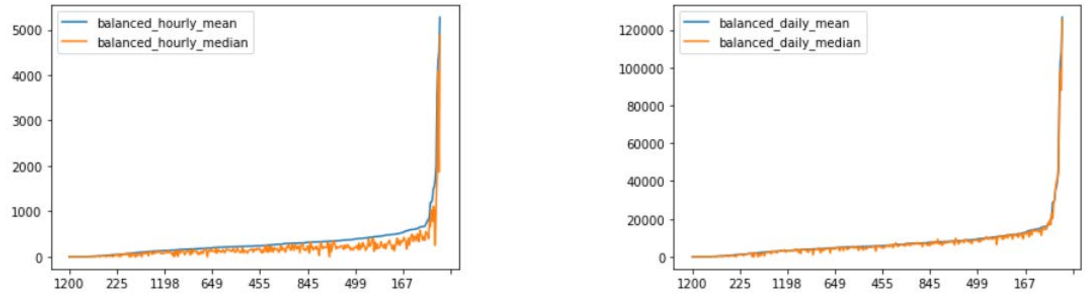
Figure 6: Median consumption of all hours within all weeks of all the data.

Comparison of mean and median consumption at hourly and daily basis

The comparison between the median and mean consumption provides additional insights, particularly on the peaks. The mean is sensitive to extreme values, while the median remains relatively unaffected. If the mean and the median are close to each other, but the median is lower than the mean, it suggests a positively skewed contribution with occasional high spikes in energy consumption. Figure 7.a shows

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

that most consumers have isolated peak profiles, as the median is often lower than the mean. If we do the same comparison for the daily mean and median (Figure 7.b) we see that the median is still lower for most consumers, but the difference is smaller than for the hourly data. That means the daily users' consumption is not as volatile as the hourly consumption.



12. Mean and median comparison at hourly level

13. Mean and median comparison at daily level

Figure 7: Comparison of the hourly mean and the hourly median

At the first glance, the increase and decrease in hourly load seems to be pretty symmetric and no sudden changes are detected (Figure 8).

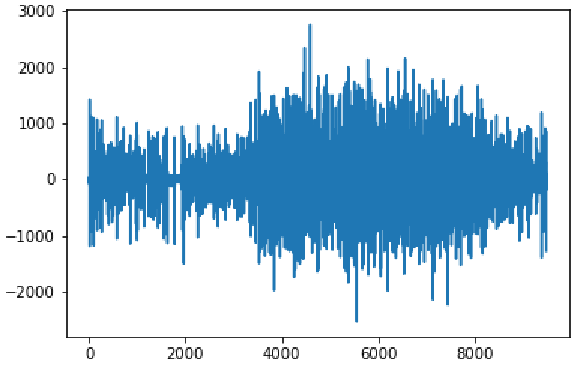


Figure 8 Changes in hourly load

Median relative increase and decrease in hourly load

The median relative increase and decrease in hourly load is calculated by first, calculating the change in load from one hour to the next, and then calculating the median from the collected decrease values. This measure is used to understand the central tendency of how much the hourly load either increased or decreased during the giving time frame (Figure 9). Figure 10 shows that some profiles have quite distinct loads, suggesting that most probably electricity was only used during a period of holidays and not during the whole year.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

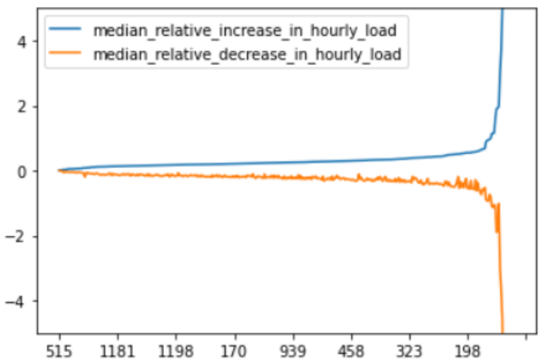


Figure 9 Median increase and decrease during hourly load

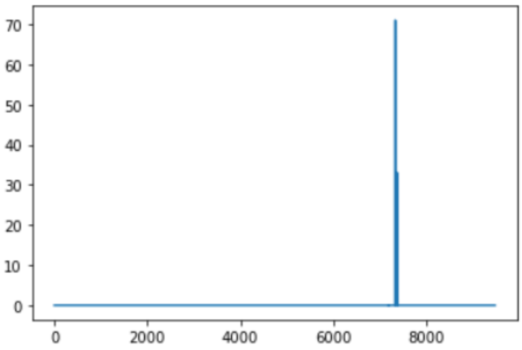


Figure 10 Load peaks most probably related to electricity consumption only during a period of holidays

2.2.2 Measure of dispersion

The measure of dispersion provides insights into the consistency or volatility of energy usage patterns over a period (eg. hourly, daily etc.). It helps understand how much the energy consumption values deviate from the average, which can be valuable for energy planning, forecasting and management.

The Pearson coefficient of variation (**CVp**) is the ration of the standard deviation to the mean.

$$CVp = \frac{std}{mean}$$

It assesses the relative variability, or dispersion, of a data set with respect to its mean, and we calculated for hourly and daily consumption for a better comparison of the two. The dispersion is more evident at the hourly level than at the daily level, suggesting that the difference of consumption is greater between hours than between days (Figure 11). The CVp for daily consumption is lower than for hourly consumption, which means that the load over days does not have as many isolated peaks as the load over hours.

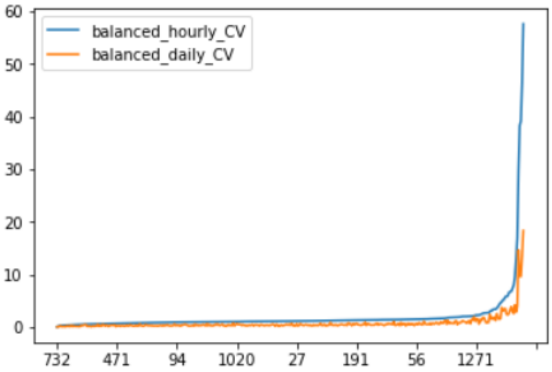


Figure 11: Mesure of dispersion at hourly and daily level. The difference of electrcity consumption is greater between hours than between days

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Pearson median skewness for hourly and daily consumption

The Pearson median skewness (Sk2) subtracts the median from the mean, multiple the difference by three and divides the product by the standard deviation.

$$Sk2 = \frac{3(mean - median)}{std}$$

It assesses the distribution with respect to symmetry. The comparison of the hourly and the daily skewness shows a different degree of skewness, meaning that the shape of the distribution between days and hours is different often with opposite signs for the consumer (Figure 12). This variation in skewness can be used as useful information to group load profiles, possibly for better understanding consumption patterns.

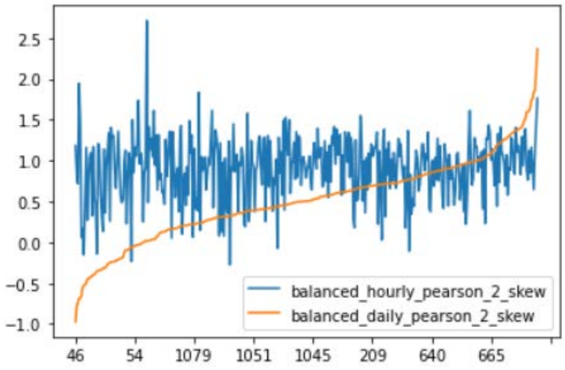


Figure 12: Pearson median skewness at hourly and daily level

Mean, median and Pearson coefficient of variation of monthly consumption

We first normalized the monthly consumption to 30 days to make months comparable. Then, we calculated the mean, the median and the Pearsons coefficient of variation for each month.

Zero consumption rate

The zero-consumption rate is the share of hours where the consumption drops to zero. Most of the occupants do not have zero-consumption because appliances, like fridges, internet routes, etc. represent constant loads. Only a few consumers have hours where their consumption drops to zero as shown in Figure 13 and those profiles can be clustered together.

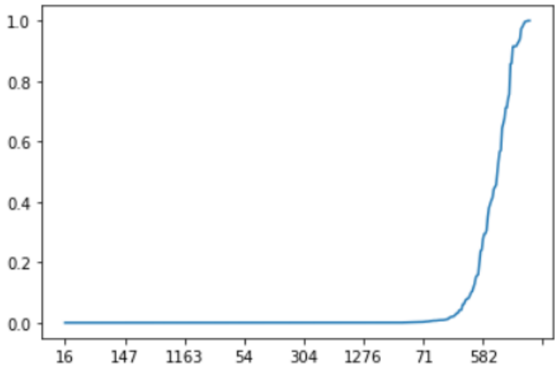


Figure 13: Zero consumption rate

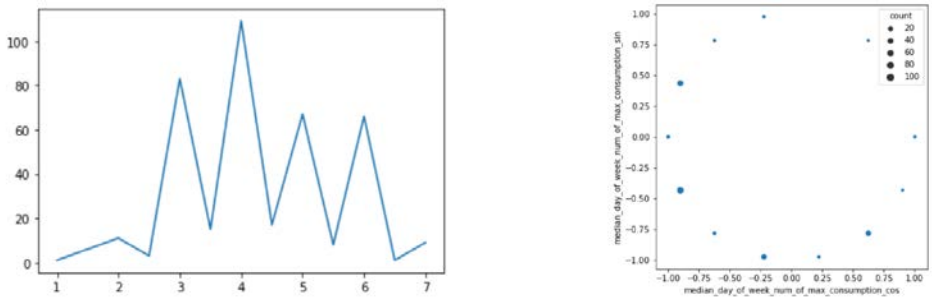
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

2.2.3 Peaks

Peak consumption-day of the week
In order to identify the day of the week with the highest consumption, we calculated the average hourly consumption for each day and then multiplied it by 24. This approach enables us to include days with less than 24 hours of available data as well.

Mean and median number of days with the highest consumption over the weeks
First, we computed the day with the peak consumption. Then, we determined the average and median values for each week, utilizing the range of days (1-7 representing Monday to Sunday).

Median day of the week with max consumption
The median day of the week with max consumption represents the initial week's day when consumption surpasses more than half of the records with maximum usage. Consequently, the days at the center hold great significance. As depicted in Figure 14.a, Tuesday witnesses most consumers utilizing less than 50% of their weekly electricity demand, whereas Thursday emerges with the highest value due to its midweek position. To capture cyclic or periodic trends in the data, this attribute undergoes circular transformation using sine and cosine functions (Figure 14.b).



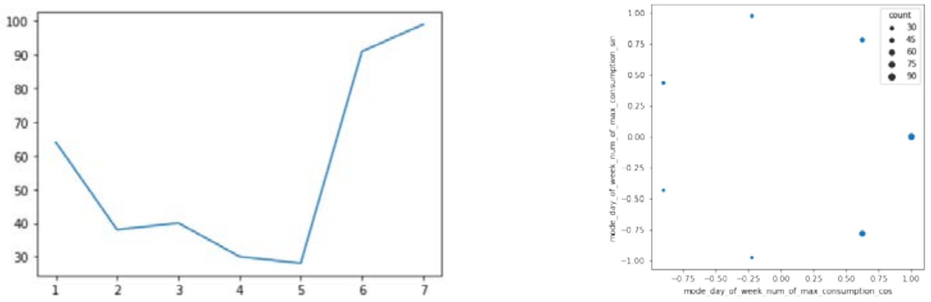
a. Day of the week with 50% of highest energy consumption

b. Circular transformation

Figure 14: Median of the days with highest energy consumption

Mode of the day of the week with the highest consumption
The mode of the day of the week with the highest consumption represents the value or values that occur most frequently in the data set. In other words, it is the value that appears with the highest frequency. This value is calculated by first determining the days of each week with the highest consumption. Then the mode of these days for all weeks is calculated. One drawback of this approach is that multimodal distributions are ignored because we only focus on the first mode. Figure 15.a shows that the highest consumption is on the weekend.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.



a. *The highest energy consumption is on the weekend*

b. Circular transformation

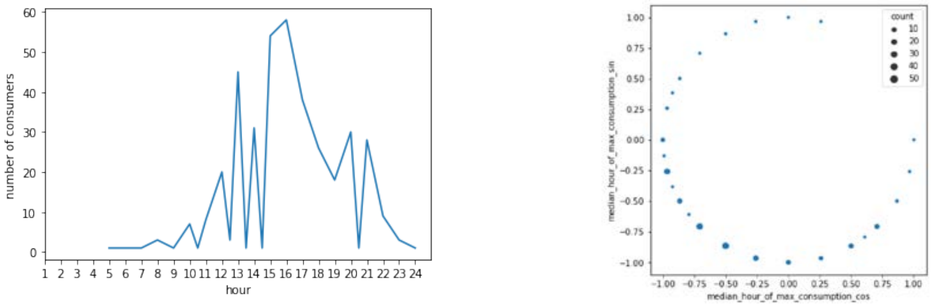
Figure 15: Mode of the weekday with the highest energy consumption

Peak consumption-hour of the day

The peak consumption-hour of the day determines the most frequently occurring hour of the day with the highest consumption. We counted when the highest consumption for the day occurs in each hour. This information is subsequently utilized by two additional features: the median and the mode hour of the day with the highest consumption.

Median and mode hour of the day with max consumption

The peak consumption hour for each day is represented by the median or the mode value and each load profile is summarized by a singular value. Figure 16.a illustrates that the according to the median, the most frequent instance falls approximately between 3-4 pm; while according to the mode the maximum consumption is between 3-11 pm (Figure 17.a).

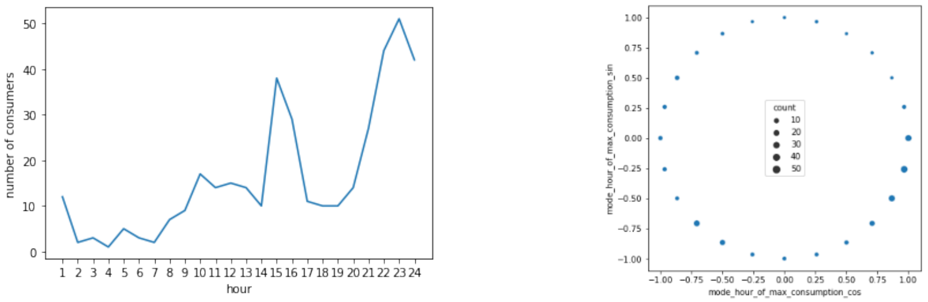


a. *The highest energy consumption is in the afternoon*

14. *Circular transformation*

Figure 16 Median hour of the day with highest energy consumption

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.



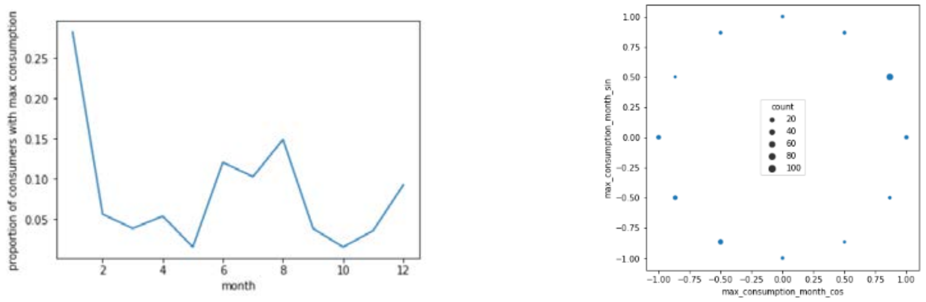
a. Highest energy consumption is in the evening

b. Circular transformation

Figure 17 Mode hour of the day with highest energy consumption

Month with maximum consumption

The month with max consumption displays the month with the highest consumption for a specific load profile. In Figure 18.a, it's evident that January has the highest consumption. Additionally, due to the cooling demand, the summer months also contribute significantly to energy consumption.



a. Months with highest energy consumption

b. Circular transformation

Figure 18 Months with the highest energy consumption

Number of slope changes per month

A "slope change" refers to the moment when the consumption pattern shifts from a decline to an upward trend or vice versa. If the monthly consumption pattern exhibits several local peaks, determining the month with the greatest consumption becomes less elucidating. As a result, we analyze the count of monthly slope transitions. In Figure 19, it is evident that most consumers experience four or six slope changes in their monthly consumption pattern over the course of a year.

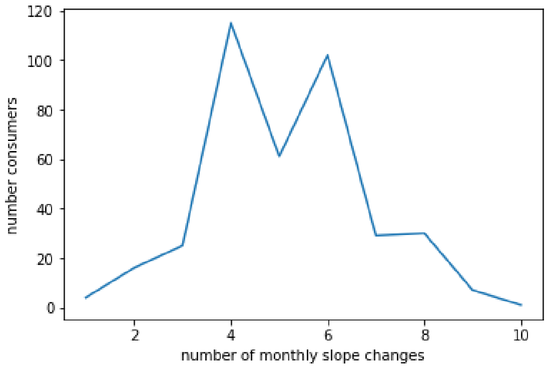


Figure 19 Slope change over a year

2.2.4 Workday vs holidays

We used a logarithmic transformation on the ratio variables to reduce the non-linear scaling impact resulting from the ratio transformations. This approach aids in achieving a more balanced and interpretable scale for the data.

Mean and median logarithmic ratio of holidays versus workdays
First, we determined the mean/median consumption for both working days and holidays (holidays include Saturday and Sunday as well as bank holidays). Then, we divided the mean/median consumption of the workdays by the mean/median consumption of the non-working days. Lastly, we apply a logarithmic transformation to the ratio variables.

$$MeanHW = \log\left(\frac{mean\ workdays}{mean\ holidays}\right) \qquad \text{MedianHW} = \log(\frac{median\ workdays}{median\ holidays})$$

By examining consumption disparities between workdays and holidays, we uncovered recurring trends, such as users who stay at home during weekends or behaviors exclusively utilized during holidays. We lastly, compared the mean and the median results to check if there is any substantial difference, as shown in Figure 20.

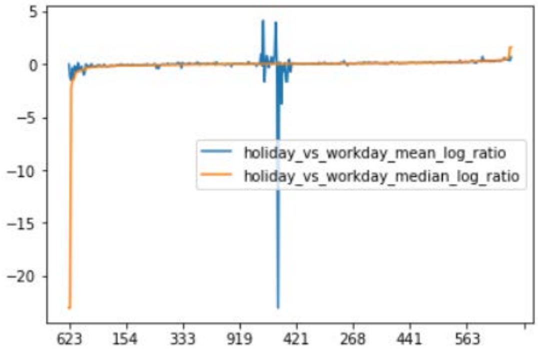


Figure 20: Comparison of the logaritmic ratio between the mean and the median of holidays vs workdays.

The logarithmic ratio between summer and winter consumption
The comparison of loads between the summer and winter seasons informs on heating systems and occupancy behavior. The ratio is calculated by dividing the energy consumption in summer by the energy consumption in winter (Figure 21). The ratio is logarithmic to prevent distortion in the distances between observations caused by ratio values below one being confined to the [0,1] interval.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

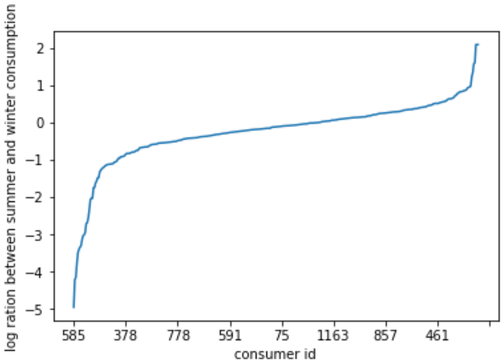


Figure 21: Logarithmic ratio of the energy consumption between summer and winter

The logarithmic ratio between workday and non-workday consumption

To catch behavior linked to holidays, we calculated the ratio of the workday to non-workdays consumption by dividing the mean workday consumption by the mean non-workday consumption. This ratio is then taken as logarithmic to prevent distortions.

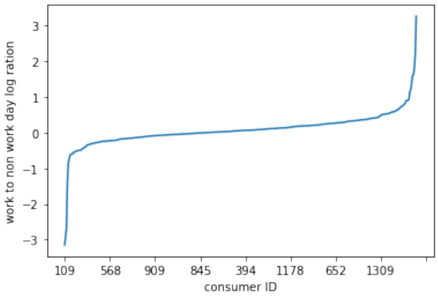
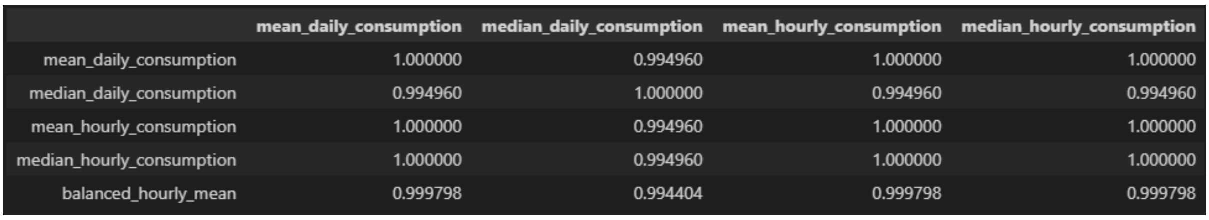


Figure 22: Logarithmic ratio of the energy consumption between workday and non-workdays

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

2.3 Features selection

After extracting new features, a comprehensive analysis was conducted to further explore the relationships and identify the most influential variables. We first performed a correlation analysis by dropping those features with a Pearson linear correlation above 0.95 (Figure 22).



	mean_daily_consumption	median_daily_consumption	mean_hourly_consumption	median_hourly_consumption
mean_daily_consumption	1.000000	0.994960	1.000000	1.000000
median_daily_consumption	0.994960	1.000000	0.994960	0.994960
mean_hourly_consumption	1.000000	0.994960	1.000000	1.000000
median_hourly_consumption	1.000000	0.994960	1.000000	1.000000
balanced_hourly_mean	0.999798	0.994404	0.999798	0.999798

Figure 23: Features with a Pearson correlation above 0,95

Our goal is to test various clustering methods with variations in their hyperparameters, different numbers of clusters (between 2 and 10), and all combinations between 2 and 10 variables (columns) from the feature dataset generated from hourly consumption data. We first tested the combination of **subsets of 2 -10 elements**, but this produced an excessive number of combinations (around 28 million feature combinations) making this approach unfeasible. So, we **grouped the features** according to the type of information that each one summarizes:

1. **Average** consumption (mean hourly/daily consumption and median hourly/daily consumption)
2. **Shape** of the consumption, such as measures of dispersion and skewness (CVp hourly/monthly consumption and hourly/daily Pearson 2nd skew)
3. **Day** of the week with the **highest** energy **consumption** for each customer (mean weekly maximum consumption, median/mode day of the week maximum consumption and sinus and cosine of the median/mode of the days with maximum consumption)
4. **Hour** with the **highest** energy **consumption** (mean hour count per day, median/mode maximum consumption, sinus and cosine of the hours with the highest consumption)
5. **Months** with the **highest** energy **consumption** (month with the highest consumption and sinus and cosine of the month with the highest consumption)
6. **Holiday consumption** (mean and median logarithmic ration between workdays and holidays, logarithmic ratio between workdays and weekends)
7. **Remaining** variables associated with diverse content (load increase/decrease, proportion of periods without consumption, changes in the slope of the series of average monthly consumptions, and the ratio of consumption between winter and summer)

We then defined all subsets of features that can be formed by taking at most one variable from each of the first groups (1-6) and any number of variables from the last group of features (group 7). This prevents us from simultaneously including more than one variable or pair of variables associated with the same type of information. Lastly, we added a condition on the number of features in each subset to be considered, which must be between 2 and 10. This reduces the number of feature combinations to be tested with different clustering strategies from around 28 million to just over 100 000.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

3 Conclusion

In this chapter, we overviewed the workflow of feature engineering to extract information on short term and long-term behavioral trends from unlabeled residential load profiles. Based on these features grouping the profiles should be facilitated.

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

4 Building stock data on different levels of aggregation

In this chapter we describe methods for handling building stock data on higher aggregated levels like neighborhoods, districts or cities, especially needed for local and regional energy planning or the operation of local energy communities or positive energy neighborhoods. By liaising with WP3 (input data) and WP5[11] (requirements from analytics/services), below aspects will be addressed in this task.

Firstly, we will develop the methodology to comprise data on the composition of different building archetypes and resulting energy consumption patterns. Secondly, we will explore how aggregated data can be split up and distributed back to the individual building level (e.g., energy consumption for the various end uses derived from consumption data aggregated on street or sector level). The latter can serve as input and/or calibration for the methods corresponding to indicator group 1 and 2. Then, we will also investigate the data requirement for assessing the spatial allocation of installed HVAC systems and methods how to deal with related data gaps. In this respect, the activity may require methods to combine (sometimes incomplete) data on building or street segment level with aggregated data on the same indicator but on a higher, aggregated level (e.g., combining hourly metering data for some specific buildings in a street with yearly energy consumptions on sector level). Herewith, special attention will be given on how to avoid data loss while ensuring that the methods and generated output remain GDPR compliant.

In this section, a literature review is primarily conducted to have an overview of the aforementioned aspects, then we selectively propose several methods for generating synthetic building stock datasets according to the provided preliminary datasets in WP3 and the service requirements in WP5.

4.1 Literature review

4.1.1 Building stock modelling and data

Building Stock Energy Modelling (BSEM) and Urban Building Energy Modelling (UBEM) are the terms often used for the building stock analysis in the reviewed literature[12]. Different methods and tools have been developed for building stock with the aim to provide various insights into system performance[13], building stock retrofitting potential[14] [15], energy driven planning[16], forecasting[17] and urban decision making[18] by evaluating factors such as short or long term energy use and demand,

[11] In WP5 different analytical tools and applications are developed to process, analyse and create additional insights from building information data.

[12] Hong T, Chen Y, Luo X, Luo N, Lee SH. Ten questions on urban building energy modeling. Building and Environment 2020;168:106508, doi:10.1016/j.buildenv.2019.106508.

[13] Allegrini J, Orehounig K, Mavromatidis G, Ruesch F, Dorer V, Evins R. A review of modelling approaches and tools for the simulation of district-scale energy systems. Renewable and Sustainable Energy Reviews 2015;52:1391–404,

[14] Hong T, Piette MA, Chen Y, Lee SH, Taylor-Lange SC, Zhang R, Sun K, Price P. Commercial Building Energy Saver: An energy retrofit analysis toolkit. Applied Energy 2015;159:298–309, doi:10.1016/j.apenergy.2015.09.002.

[15] Kristensen MH. Urban building energy modelling for retrofit analysis under uncertainty, Aarhus Universiet; 2018.

[16] Ferrari S, Zagarella F, Caputo P, Bonomolo M. Assessment of tools for urban energy planning. Energy 2019;176:544–51, doi:10.1016/j.energy.2019.04.054.

[17] Hu Y, Cheng X, Wang S, Chen J, Zhao T, Dai E. Times series forecasting for urban building energy consumption based on graph convolutional network. Applied Energy 2022;307:118231, doi:10.1016/j.apenergy.2021.118231.

[18] Kavgic M, Mavrogianni A, Mumovic D, Summerfield A, Stevanovic Z, Djurovic-Petrovic M. A review of bottom-up building stock models for energy consumption in the residential sector. Building and Environment 2010;45:1683–97, doi:10.1016/j.buildenv.2010.01.021.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

short term demand response, GHG emissions, potential renewable energy generation and storage etc. These approaches are generally classified to bottom-up and top-down approaches.[19]

The top-down approach provides insights for forecasting and long-term planning and policy evaluation using macro-economic and statistical data[20]. Top-down models rely on macro level information, historical data and statistical energy use, socio-economic factors and energy prices to estimate energy consumption or carbon dioxide emissions for long term purposes such as high-level building energy policy evaluation, while bottom-up models start from detailed individual building level data and scale all the way up to street, neighborhood, district, city, regional and national building stock level[21]. Top - down and bottom-up approaches are further subdivided into constituent sub-categories by different researchers. For instance, Ali et al[22] further divide bottom-up approaches into three main sub-groups: physics-based, data driven and reduced ordered methods. In IEA Annex 70[23], Langevin et al.[24] propose a multi-layer quadrant scheme that classifies modeling techniques by their design (top-down or bottom-up) and degree of model transparency (black-box or white-box). In general, bottom-up approaches become more and more popular due to the emerging data at building level. Specifically, a hybrid approach in-between top-down and bottom-up can be favorable in terms of applicability.
Among bottom-up approaches, physics-based methods have the advantage that they enable the assessment and quantification of the combined effect of several technologies on the building energy demand, and do not require detailed historical energy consumption and socio-economic factors[25].

Input data quality have a crucial impact on the performance of analysis such as UBEM or BSEM. Hence, data enhancement techniques play an essential role in improving the raw data source quality and forming a reliable synthetic dataset that can sufficiently represent the targeted building stock and be further used in the dedicated building stock analysis. In principle, to generate such synthetic building stock datasets, data sources are clustered to three groups by Nägeli et al[26]: data on building stock structure and spatial distribution (e.g. geo-referenced data on number of buildings), data on building stock characteristics (e.g. u-value of building components) and data on building usage (e.g. energy consumption, number of occupants). Depending on specific research topic and data availability, different methods can be applied to generate the required synthetic datasets. These are reviewed and selectively presented in the following sections.

[19] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable and Sustainable Energy Reviews 2009;13:1819–35, doi:10.1016/j.rser.2008.09.033.

[20] Reinhart CF, Cerezo Davila C. Urban building energy modeling – A review of a nascent field. Building and Environment 2016;97:196–202, doi:10.1016/j.buildenv.2015.12.001.

[21] Li W, Zhou Y, Cetin K, Eom J, Wang Y, Chen G, Zhang X. Modeling urban building energy use: A review of modeling approaches and procedures. Energy 2017;141:2445–57, doi:10.1016/j.energy.2017.11.071.

[22] Ali U, Shamsi MH, Hoare C, Mangina E, O'Donnell J. Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis. Energy and Buildings 2021;246:111073, doi:10.1016/j.enbuild.2021.111073.

[23] https://energyepidemiology.org/

[24] J. Langevin, J.L. Reyna, S. Ebrahimigharehbaghi, N. Sandberg, P. Fennell, C. Nägeli, J. Laverge, M. Delghust, É. Mata, M. Van Hove, J. Webster, F. Federico, M. Jakob, C. Camarasa, Developing a common approach for classifying building stock energy models, Renewable and Sustainable Energy Reviews, Volume 133, 2020, 110276, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2020.110276.

[25] https://doi.org/10.1016/j.energy.2017.11.071

26 Nägeli, C.; Thuvander, L.; Wallbaum, H.; Cachia, R.; Stortecky, S.; Hainoun, A. Methodologies for Synthetic Spatial Building Stock Modelling: Data-Availability-Adapted Approaches for the Spatial Analysis of Building Stock Energy Demand. Energies 2022, 15, 6738. https://doi.org/10.3390/en15186738

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

4.1.2 Archetype modelling and resulted energy consumption patterns

Based on the general overview described in Section 3.1, this section focuses on bottom-up, physical-based techniques, given its flexibility with regards to data availability. This type of urban energy modeling requires the definition of model data inputs regarding the modeled buildings' geometry, construction assemblies, HVAC systems and usage patterns, as well as climate conditions. However, such detailed data collection efforts become impractical for larger urban areas and the computational efforts become often too excessive in case one has to setup an individual building model for each building in large urban areas. Therefore, different approaches have been developed to reduce the modeling and simulation efforts, of which most can be classified as one of three main approaches:

1. *Distribution approach*[27]
   This approach determines the end-use energy consumption from regional or national distributions of appliance ownership and use. Although this method relies on national figures of appliance penetration and may adopt historic energy consumption, they are classified as bottom-up due their end-use disaggregation.

2. *Sample approach*[26]
   In this technique, actual sample building data is used as the model input information, after which the total building stock energy consumption can be estimated by applying appropriate weights to the results. However, this method requires an extensive database to represent the entire modelled building stock in case of large or highly diverse regions.

3. *Archetype-based approach*
   This widely-used technique classifies the building stock according to several building characteristics, after which the energy consumption estimates of modeled archetypes are scaled up to be representative of modeled housing stock.

Among these approaches, we believe that archetype modelling is the most generally applicable method because of its flexible data requirements, has a large potential for improving UBEM [28], and has widely been accepted by both academia and practitioners. Therefore, we focus on archetype modeling in the rest of this section.

In archetype modeling, abstraction of the modeled building stock is made into "building archetypes", i.e. sets of either representative samples or artificially created buildings that characterize subsets of buildings with similar properties[29]. The number of buildings in the modeled area that correspond to each building archetype is then estimated based on national census statistics or available data of the archetype indicators for each individual building. Each building archetype is then modelled using a chosen simulation engine to estimate its energy consumption, after which these estimates are further scaled up to represent the regional or national building stock through aggregation[30].

Dahlström et al.[28] describes that setting up a UBEM framework based on archetypes comprises five distinct main processes which follow each other chronologically: (1) data acquisition and processing, (2) building stock segmentation and archetype development, (3) simulation (modelling components), (4) model calibration, and (5) model application.

[27] https://www.sciencedirect.com/science/article/pii/S0301421507003291
[28] https://doi.org/10.1016/j.enbuild.2022.112099
[29] https://doi.org/10.1016/j.energy.2019.04.197
[30] https://www.sciencedirect.com/science/article/pii/S1364032108001949

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

In this section, we focus on the archetype development, which comprises three major steps, namely, segmentation, characterization and calibration, and quantification [28] [31] [32]:

1. *classification or segmentation: buildings are grouped according to one or more indicators,*
2. *characterization: each archetype has to be characterized by a complete set of thermal and building physics characteristics, including construction materials, usage patterns, and building systems, and*
   calibration: calibration and validation of uncertain archetype parameters,
3. *quantification: determine the number of buildings belonging to each building archetype.*

4.1.2.1 Archetype segmentation/ classification

In this first step, buildings are grouped according to one or more indicators or criteria, which need to be 1) correlated to the energy demand of the building, and 2) available for all buildings.

As described by Dahlström et al.[28], studies involving archetype classification have adopted different methodologies depending on the aim of the study. First, considering the input data that is used for the classification process, Dahlström et al.[28] divided studies based on whether or not they have utilized any variant of an EPC database. Herewith, EPCs can be considered as a highly value data source that can eliminate some of the data gathering or reliability issues compared to studies using other data sources.

Second, segmentation schemes to split the building stock can either be defined by the modeler in a manual (deterministic), semi-automatic (statistic) or fully automatized (data-driven) way. For the latter, the development of machine learning techniques (both supervised and unsupervised) allows for more automated statistic segmentation, i.e. clustering, where the modeler's main input is to define the building (energy) similarity metrics to be used, and not the feature splits themselves [33] [34] [35].

Typical indicators involve socio-economic[36] [37] [38](type of building usage, income level), spatial (climate zone, location) [39], structural [40] [38] (e.g. age, floor area, envelope form, number of floors), energy installation[41] (e.g. heating source, ventilation system, status of refurbishment) or performance (e.g. energy use intensity, total energy, peak power) features[42] [43].

The indicators most often used to classify buildings into archetypes are programmatic use (e.g. residential, office, retail, etc.), floor area, shape typology and age of the construction.

[31] https://www.sciencedirect.com/science/article/pii/S0378778819306553
[32] https://www.sciencedirect.com/science/article/pii/S0360132314001991
[33] https://doi.org/10.1016/j.enbuild.2019.109364
[34] https://doi.org/10.1016/j.enbuild.2012.03.033
[35] https://doi.org/10.1007/978-981-19-1280-1_14
[36] https://doi.org/10.1007/s12053-017-9609-1
[37] https://doi.org/10.1016/j.egypro.2017.03.018
[38] https://doi.org/10.1016/j.enbuild.2021.111175
[39] https://doi.org/10.1016/j.enbuild.2017.08.029
[40] https://doi.org/10.1016/j.enpol.2014.01.027
[41] https://doi.org/10.1016/j.enbuild.2018.08.032
[42] https://doi.org/10.1016/j.egypro.2017.03.244
[43] https://doi.org/10.1016/j.energy.2018.05.190

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

The number of archetypes varies significantly between different studies. Monteiro et al.[44] concluded that the modelling accuracy increases with the number of archetypes, although an increased number of archetypes increase the data requirements and increases the computational efforts. Therefore, a compromise needs to be made and the adopted number of archetypes depends on the diversity of the modeled building stock, the building energy metric of interest, the data availability for segmentation and characterization and the requirements for accuracy and computational efforts.

4.1.2.2 Archetype characterization

The second step in the archetype development involves the characterization of the identified building categories for all relevant energy simulation parameters. Besides the building geometry, these include all non-geometric building and occupant factors which influence energy demand, including envelope construction details, HVAC system properties, occupancy schedules, internal loads, etc. The exact set of parameters to be defined depends on the UBEM simulation tool, the thermal modelling approach (steady state versus dynamic) and the model zoning simplification (single zone vs multi zone).

In general, archetype characterization can either be done deterministic [38] [44], i.e. a single value assigned to each parameter and used for every building, or probabilistic by defining parameters as distributions [39] [45]. Especially for parameters characterized by high uncertainties, a probabilistic characterization allows to have a more reliable estimate of the resulting energy demand. These typically involve parameters related to occupant behavior and preferences, but also any parameter that is often not found in audit, survey, or GIS data[46]. The latter can include, for example, infiltration air exchange rates, which are difficult to measure, thermal losses from HVAC distribution systems, or the amount of unconditioned floor area in the building.

The adopted values for the energy simulation parameters can be derived from building data, expert knowledge, literature and building surveys. Absent data or data of insufficient granularity can lead to oversimplified and biased archetype characterization, in which case various calibration methods should be applied (see next section).

4.1.2.3 Calibration

Validation data and calibration data are crucial for UBEM, as validation data can be employed to evaluate the performance of UBEM, and calibration data can be used as a benchmark to adjust the input parameters in UBEM. Calibration methodologies for building energy models are being used more and more in the literature, of which the probabilistic calibration approaches, e.g. approaches based on Bayesian inference, have become increasingly popular in recent years[47].

[44] https://doi.org/10.1016/j.egypro.2017.03.244
[45] https://doi.org/10.1080/10789669.2011.582920
[46] https://doi.org/10.1016/j.enbuild.2016.10.050
[47] https://doi.org/10.1080/19401493.2012.723750,
https://doi.org/10.1016/j.enbuild.2016.04.025,
https://doi.org/10.1016/j.buildenv.2014.12.016,
https://doi.org/10.1007/s12273-016-0291-6,
https://doi.org/10.1016/j.enbuild.2017.08.069,
https://doi.org/10.1016/j.enbuild.2016.10.050,
https://doi.org/10.1016/j.enbuild.2017.08.029,
http://www.ibpsa.org/proceedings/BS2015/p2435.pdf

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Optimization and Bayesian calibration are both approaches to choosing vectors from the input parameter space that result in the lowest calibration error and involve the minimization of an objective function (i.e., calibration error) by changing the model's input parameters.

However, primarily limitations in data access, time and computational power made it difficult to apply these methods for urban building energy models. Since measured energy data for individual buildings is rarely available in UBEM, often only a single value (a district's annual energy consumption) can be used for model validation. For this purpose, Bayesian calibration is recently being used [48] [49] to address uncertainty of individual building parameters by characterizing each parameter undergoing calibration as a probability distribution instead of a single value, and subsequently using measured data points to update these to posterior distributions. Herewith, the combination of UBEM's with surrogate regression models enables to reduce the often high simulation times [50].

4.1.2.4 Archetype quantification

The quantification step determines the distribution of archetype buildings in order to be representative of the building stock, i.e. the simulation results of each archetype are weighted by the number of buildings of the modeled building stock corresponding to the archetype.

To quantify the number of building corresponding to each archetype and compute their total floor area, studies on larger scale typically use national statistics. For studies on smaller scale (e.g. neighborhood level), more detailed data with regard to the archetype indicators could be collected for the modeled building stock, allowing to compute the number and summed floor area for each archetype based on the more detailed data.

4.1.3 (dis)aggregations

The process of disaggregation of energy data known at an aggregated level, e.g. city or national level, to individual building level has been studied extensively in the literature. Moreover, it is closely linked with synthetic building stock energy modelling (SBSEM)[51], which allows to model spatially distributed synthetic building stocks. In this review, we first give present promising methodologies for SBSEM that are useful for building energy data disaggregation. Subsequently, we discuss two other types of techniques used in the literature that allow to predict the individual energy consumption of buildings based on data on an aggregated level, and as such downscale the city or neighborhood energy use to the level of buildings.

4.1.3.1 Synthetic building stock energy modelling

Synthetic building stock energy modelling (SBSEM) refers to the field of generating disaggregated data of individual buildings in building stocks based on aggregate data. Nägeli et al.[26] used two approaches of using SBSEM, the so-called sample-based and sample-free SBSEM. The sample-based approach is based on the Iterative Proportional Updating (IPU) approach[52] and relies on the use of a sample dataset of individual buildings. Using a standard Iterative Proportional Fitting (IPF) procedure[52], they spatially distribute a sample set of building records to match an aggregated dataset describing the spatial distribution of the building stock. The sample-free approach reconstructs the synthetic building

[48] https://doi.org/10.1080/19401493.2020.1729862
[49] https://doi.org/10.1080/19401493.2012.723750
[50] https://doi.org/10.1080/19401493.2018.1457722
[51] https://doi.org/10.1016/j.enbuild.2018.05.055
[52] Ye, Xin & Konduri, Karthik & Pendyala, Ram & Sana, Bhargava & Waddell, Paul. (2009). Methodology to match distributions of both household and person attributes in generation of synthetic populations.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

stock based on aggregate data describing the structure and spatial distribution of the building stock. Herewith, the synthetic building stock is initialized one by one by iteratively generating buildings and assigning different building characteristics based on the probabilities and constraints in the composition of the building stock of the modelled area. In both approaches, the synthetically created building stock is then enriched with additional attributes needed for energy modelling by stochastically assigning attributes based on distributions or assigning data based on archetype data.

4.1.3.2 Energy data disaggregation

As explained earlier, bottom-up statistical techniques are used to derive relationships or correlations between key input parameters and output parameters such as whole building or end use energy consumption[53]. Mastrucci et al.[54] and Howard et al.[55] used (multiple) linear regression models to downscale measured natural gas and electricity consumption from the aggregated zip code level to single dwellings, and apportioned these consumptions to the different end-uses.

Zhang et al.[56] used a combination of statistical matching and various machine learning techniques (linear regression, gradient boosting regression, and random forest regression, Support Vector Machine, …) to enrich a sample dataset with energy data, and subsequently use the IPU algorithm to generate a synthetic population of households for the entire metropolitan region. Similarly, Robinson et al.[57] employed multiple machine learning methods to estimate the commercial building energy consumption in diverse metropolitan areas in the United States.

4.1.4 Spatial allocation of HVAC systems in UBEM

This part of the literature review investigates the spatial allocation of HVAC systems in an urban energy system through data analysis techniques. We first mention how the HVAC information can be devised in a building dataset. Then, we explain how the lack of minimum requirements can be solved according to the literature. We hereby distinguish between spatial allocation of HVAC and thermal load prediction of buildings. Although estimation of heating and cooling loads based on open data has been widely explored, few studies in the literature have focused on the spatial allocation of HVAC systems using open data.

HVAC systems in UBEM can be defined based on their metadata description (e.g. fuel type, type of system, emission system), although usually, only the system's efficiency is sufficient to investigate economic and environmental indicators under different scenarios. The minimum required information about the installed HVAC depends on the use-case. The level of detail about the HVAC for detailed and case specific energy simulations is not like the large-scale, such as building stock, energy simulations. The simplest technique is to overcome the lack of required data is to estimate the details of the installed HVAC for each building in the dataset based on its archetype and/or the energy consumption patterns (if available). For example, a newly built house can be assumed to have a mechanical ventilation system. With similar reasoning a significantly higher gas consumption for a building, street or district in winter may indicate the dominance of gas boilers. However, a variety of methods have been used to deal with lack of data about the HVAC and to reach different goals as exemplified below:

[53] https://doi.org/10.1016/j.rser.2020.110276
[54] https://doi.org/10.1016/j.enbuild.2014.02.032
[55] https://doi.org/10.1016/j.enbuild.2011.10.061
[56] https://doi.org/10.1016/j.energy.2018.04.161
[57] https://doi.org/10.1016/j.apenergy.2017.09.060

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

a. Financial analysis for energy system design for retrofit planning or energy system design using an integrate tool (e.g. CityBES [58]). This tool takes the HVAC allocation as input and adds it to its original dataset. This is the most common approach for spatial allocation of HVACs in UBEM.

b. Energy storage feasibility study e.g., Chambers et al. [59].

In this case, the study does not explore the existing HVAC but the feasibility of a district heating system with storage capacity using a projection for upcoming HVAC installations. In this case, the HVAC type for each building is imposed by the modeler to be district heating.

c. Energy flexibility investigation, e.g FlexiGIS open source tool [60].

In this type of studies, a tool is developed that is using electricity and gas consumption time series as inputs and do not simulate HVAC. As such, spatial allocation of HVAC is done indirectly via the energy consumption data.

d. Carbon emission reduction potential [61].

In this example, a comprehensive scenario analysis was carried out. The modeling included white-, black-, and grey-box approaches depending on the part of the system and it requires a significant amount of data to build up the model. HVAC spatial allocation is done outside the energy simulation model in a separate optimization model, of which the output (selected HVAC) serves as input for the energy model. The selected HVAC is modeled with white-box models including a model of undefloor heating, air conditioning etc. One of the targets is to devise the clusters of buildings that could operate with similar supply temperature for the emission system. Then, the feasibility of a district heating/cooling system can be compared to that of discretized HVACs.

e. Fault detection (e.g. Buffa et al. [62]).

In the framework of the H2020 project RELaTED [63], two tools have been developed for automatic fault detection in DH substations based on ML algorithms: DH doctor and DH Autotune [64]. The first one exploits clustering, and it is based on daily averaged readings. Both methods rely on data-driven approaches to identify HVAC operation and hence can be used to identify HVAC type. Herewith, HVAC identification can be considered a pre-processing step

[58] Chen, Y., Hong, T., & Piette, M. A. (2017). Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. Applied Energy, 205, 323–335. https://doi.org/10.1016/J.APENERGY.2017.07.128

[59] Chambers, J., Zuberi, S., Jibran, M., Narula, K., & Patel, M. K. (2020). Spatiotemporal analysis of industrial excess heat supply for district heat networks in Switzerland. Energy, 192, 116705. https://doi.org/10.1016/J.ENERGY.2019.116705

[60] Alhamwi, A., Medjroubi, W., Vogt, T., & Agert, C. (2019). Development of a GIS-based platform for the allocation and optimisation of distributed storage in urban energy systems. Applied Energy, 251, 113360. https://doi.org/10.1016/J.APENERGY.2019.113360

[61] Fonseca, J. A., & Schlueter, A. (2015). Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. Applied Energy, 142, 247–265. https://doi.org/10.1016/j.apenergy.2014.12.068

[62] Buffa, S., Fouladfar, M. H., Franchini, G., Lozano Gabarre, I., & Andrés Chicote, M. (2021). Advanced Control and Fault Detection Strategies for District Heating and Cooling Systems—A Review. Applied Sciences, 11(1), 455. https://doi.org/10.3390/app11010455.

[63] http://www.relatedproject.eu

[64] http://www.relatedproject.eu/wp-content/uploads/2019/10/RELaTED_D2_4_Energy_Flexibility_and_DH_Control_V4.0.pdf

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

for input data. Using smart meter data for example, one can first identify the installed HVAC and afterwards use the metered data for additional analysis, calibration and/or validation.

Integrated urban building energy models, e.g. CityBES [65] and TEASER [66], rely on multiple layers of data to make an integrated dataset for integrated models and simulations. These models have their specific requirements for input dataset including HVAC spatial allocation. Each building is specified with a tag and contains a variety of input data including its installed HVAC. As such, spatial allocation of HVAC is a pre-processing step to these integrated models. If the installed HVAC for a building is not known, other methods must be deployed to define the HVAC for a building. There are methodologies for pre-processing the building data to define its installed HVAC. For example, Yu et al. [67] developed an algorithm for online operational signatures of HVAC systems and examine their energy profiles, which could also be used to label the installed HVAC in a building. However, existing methods focus on the analysis of individual buildings, thus upscaling the applications and implementation of these methods is needed to be useful for UBEM.

Another related scientific trend is modeling electricity grid using open source data, which are often data-driven. Many recent scientific modeling projects like GENESYS [68], SciGRID [69], open_eGo [70] openMod [71] and Open Power System Data [72] and OpenGridMap focus on open source software and/or open grid data. Moreover, several grid simulation software packages exists, such as PyPSA [73] or osmTGmod [74]. To our knowledge, such models have not been employed for UBEM yet, while they may provide interesting opportunities. For instance, the data can be analyzed in order to find out the pattern in electricity use and allocate the installed HVAC to a specific location. Chen et al. [75] use smart meter data analysis to estimate the penetration of heat pumps in a city. They developed a classification method to characterize air conditioning penetration patterns with spatiotemporal resolution. Miller et al. [76] provide an overview of unsupervised data mining techniques to classify the HVAC operational indicators. Such techniques can be used to classify the installed HVAC in a building.

[65] Agbonaye, O., Keatley, P., Huang, Y., Ademulegun, O. O., & Hewitt, N. (2021). Mapping demand flexibility: A spatio-temporal assessment of flexibility needs, opportunities and response potential. Applied Energy, 295, 117015. https://doi.org/10.1016/J.APENERGY.2021.117015.

[66] Remmen, P., Lauster, M., Mans, M., Fuchs, M., Osterhage, T., & Müller, D. (2018). TEASER: an open tool for urban energy modelling of building stocks. Journal of Building Performance Simulation, 11(1), 84–98. https://doi.org/10.1080/19401493.2017.1283539

[67] Yu, X., Ergan, S., & Dedemen, G. (2019). A data-driven approach to extract operational signatures of HVAC systems and analyze impact on electricity consumption. Applied Energy, 253, 113497. https://doi.org/10.1016/J.APENERGY.2019.113497

[68] https://www.vde-verlag.de/proceedings-de/453550041.html
[69] http://scigrid.de
[70] https://www.next-energy.de
[71] https://github.com/rwl/PYPOWER
[72] https://open-power-system-data.org
[73] http://http://pypsa.org
[74] https://github.com/wupperinst/osmTGmod
[75] Chen, M., Sanders, K. T., & Ban-Weiss, G. A. (2019). A new method utilizing smart meter data for identifying the existence of air conditioning in residential homes. Environmental Research Letters, 14(9), 094004. https://doi.org/10.1088/1748-9326/ab35a8

[76] Miller, C., Nagy, Z., & Schlueter, A. (2018). A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renewable and Sustainable Energy Reviews, 81, 1365–1377. https://doi.org/10.1016/J.RSER.2017.05.124

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

To conclude, there is not a single, optimal technique for spatial allocation of HVACs in a UBEM and the optimal strategy depends on the data availability. However, three main approaches can be listed:

- *In some cases, the input building dataset already includes information (e.g. EPC certificate) about the installed HVAC systems.*
- *Classification techniques can be used to identify the installed HVAC for buildings for spatial allocation of HVACs in urban scale as input to the model data if smart meter data for gas and electricity are accessible.*
- *Correlation between input data can also be used as a pre-processing step that facilitates data enhancement for spatial allocation of HVAC. For example, known or estimated building physical parameters, insulation level, annual gas and electricity consumptions, etc. can give a good indication about the installed HVAC.*

4.2 Proposed methods based on available datasets

According to the literature review in section 3.1, UBEM necessitates certain input data. As a result, a pre-processing step is required to enhance the quantity and quality of the input data so that they meet the model requirements. The data enhancement method, on the other hand, is heavily dependent on the characteristics of the original input dataset. In this section, we first mention data enhancement procedures related to variety of datasets. Then, the datasets used in this work package are described, after which the best data enhancement method for each dataset is provided.

Based on current analysis of the datasets available in WP3 and the preliminary requirements of services/tools in WP5, following methods are proposed:

- *Data-driven archetypes are defined based on commonly available data sources (e.g. geometrical data and construction year) in EPC database, linking with the consumption profiles (both annual and high frequency profiles when data is available). This method enables a reliable calibration of the urban energy model. Elaborate datasets which contain building parameters and monitored data such as energy use and indoor temperature are useful in this method. Individual buildings of the elaborate dataset are found among the statistically defined archetypes. Then, the discrepancies between model outcomes and monitored data are studied. For instance, shares of modelling errors and archetype misclassification are studied in this analysis.*
- *Building characteristics in the database (e.g. u-value, HVAC systems) can be enriched based on construction year or other variables that define the archetypes (sampling method). EPC databases can help in deriving correlations between variety of building parameters. Monitored energy use or energy bills can be useful in this step only if they can be related to building attributes.*
- *Detailed energy use profiles can be derived from pre-processing of metadata. The aggregated energy use of a building may not suffice for planning and decision for an district/urban energy system. Detailed energy profiles can help in deriving the energy breakdown of the building, hence helping in decision making and energy planning. Energy breakdown describes the energy use of a building disaggregated for a variety of services in the building. Energy bills cannot solely suffice in this method. Thus, pre-processing the dataset can help show where and how the energy is consumed in a building. Then, the energy planer can propose tailored measures to improve energy performance of each building in a district.*

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

- *Disaggregation of national/regional data using top-down approach to breakdown national/city energy balance values (or other stats) to building (archetype) level. First, the data from a national or regional level database are analysed. Disaggregation of the data is then performed based on additional data from other resources. For example, the distribution of the building parameters in the region of the reported original database is studied. Archetypes are then defined such that the entire region is covered (can be only data-driven). Buildings of the original dataset are characterised based on the archetypes. Spatial data are added to the database if available. As such, the national/regional dataset can be disaggregated using synthetic datasets.*
- *National/regional EPC datasets are used to allocate HVAC systems to individuals, together with more detailed data when available. Installed HVACs in buildings are classified based on multiple parameters such as year of installation, energy (gas and electricity) consumption, etc.*
- *HVAC systems and design settings are identified based on metered data on the individual level. Smart meter data can reveal if a heat pump or gas boiler is devised as production unit. Year of construction, renovation year, geometrical data etc. can help define the installed HVAC. This entails deploying probabilistic methods for judgment based on historical data of the region of the study.*

A preliminary overview on the description of the available datasets and required data enhancement method are presented below based on the private/public data sources provided by seven partners in WP3 and according to the provided literature review.

15. *ENERCOOP: annual electric load profiles of more than 1000 buildings in Spain*
    As the time series are anonymized, data-driven models can be trained with this dataset. The models can produce synthetic time series on individual building level or on district level. Although this data cannot be directly related to building parameters, they can be used for data calibration.

16. *IVE: EPC database in Valencian region, Spain.*
    This dataset provides spatial data on the individual building level with the estimated annual CO_2 emission and energy demand for each building, and thus distribution approach is relevant to this dataset.

17. *Synvasion: monitoring data of 2 buildings located in Hannover, Germany*
    This dataset includes elaborate data which will be of use for model calibration.

18. *Würth: monitoring data of 12 retail shops located in Italy*
    Given that the data is for shops, this dataset can be used for archetype quantification using data-driven approaches. The correlation between weather data and the monitored data can also enhance the dataset.

19. *Veolia: 2 buildings with monitored data (energy consumption - space heating, DHW, and temperature, 15 minute resolution) and building characteristics info (floor area, number of floors, and insulation)*
    This dataset includes elaborate data which will be of use for model calibration.

20. *Köhler and Meinzer: energy (electricity bills) of single flats (anonymized) and survey data of another 68 buildings. Archetype classification can be performed using this dataset. Moreover, calibration of the models can be conducted based on the energy bills in this dataset.*

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

21. *VITO: annual gas and electricity consumption profiles of 100 Flemish households; Street level consumption data per energy (electricity/gas), injection/purchase and per main municipality at street level; Building geometrical data in Flanders; EPB (Energy Performance and Indoor Climate, new built or renovation) database in Flanders.*

This dataset requires a multi-step combined method for data enhancement. This includes disaggregation of data using probabilistic methods, characterization of buildings for archetype modelling, sampling methods for calibration of urban building energy modelling.

In addition, we also include EPC datasets of different countries (where available) as an additional data source for the building stock analysis. A preliminary exploration of open source EPC data in the EU (UK, NL, FR, BE-3 regions, IT-Lombardia region)[77] was conducted to seek the synergies, so that we can further select the most generic data enhancement method accordingly.

Besides the available datasets, past expertise and knowledge will influence the choice for the proposed methodology(ies). For instance, VITO has previously developed the Urban Energy Path Finder (UEP)[78], which is a decision support tool for future scenario analysis for energy planning purposes. It provides a holistic energy solution by calculating energy, CO_2 savings, and financial conditions for renovation scenarios and energy technology measures at building, district and city level. These scenarios include a mix of technological measures such as district heating/cooling networks, building renovation measures and decentralized renewable energy production technologies.

The workflow of UEP can be subdivided into three main parts:

a. *Characterisation of the existing situation of the buildings in the modeled district.*
b. *The evaluation of renovation measures on individual building level.*
c. *The evaluation of district heating potential for the district.*

The focus of this task is closely linked with the first step in UEP, which will be elaborated more in detail below.

UEP first gathers all available information on individual building, including the building geometry, construction year, building function, installed HVAC, etc. Next, primary characteristics and input data such as current actual energy consumption data are where needed transferred, processed, re-calculated from higher aggregation levels towards the building level by means of spatial allocation algorithms.

Next, UEP employs a bottom-up, archetype-based approach in which buildings are classified based on their function, type (apartment, terraced, semi-detached, detached) and construction year period. The archetype characterisation was done probabilistic and derived from the national EPC database. Herewith, for each archetype, each energy-related building parameter is described by a probability distribution. To determine an input value needed for the actual building energy simulations, a single value is sampled from these distributions for each parameter and each building. Since two buildings of the same archetype will be characterized by different sampled values, this methodology takes into account the natural spread of the modeled building stock.

Eventually, we will investigate the suitable methodologies that could be implemented to the project datasets and explore the possibilities to generalize these methodologies to other open source or private datasets.

[77] Arcipowska, A., Anagnostopoulos, F., Mariottini, F., Kunkel, S., Rapf, O., Atanasiu, B.,& Dumitru, M. (2014). A Mapping of National Approaches Energy Performance Certificates across the EU BPIE Review and Editing Team.
[78] https://www.energyville.be/en/research/urban-energy-pathfinder

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

5 Link to other work packages

The methodologies described in this deliverable are relevant for the following tasks in this WP. Each of these tasks gets data collected in WP3 as input together with additional open source datasets. Further, in each task a methodology and application are developed that can then be further used in WP5 and integrated into the MODERATE Platform. In the following paragraph we shortly describe the data flows in between these three work packages at our current understanding. As the project continues the exact linkages especially between WP4 and WP5 will be further specified and might change.

Figure 29 visualizes the interlinkage of WP4 with WP3 and WP5. Task 4.2[79] aims to generate synthetic data of energy consumption in buildings (electricity for lighting and appliances, space heating, domestic hot water and air conditioning) on a seasonal or yearly basis. Therefore, the input data potentially includes population data, building archetypes of a certain region and historical energy usage or costs. The output format of Task 4.2 (purple) are single values with metadata description. This data will be further used in WP5 in the Subtasks 5.1.2 (Energy system optimization and Fault Detection and Forecasting), 5.1.3 (M&V for Building energy assessment and the Benchmarking tool) and 5.1.4 (EPC Harmonization).

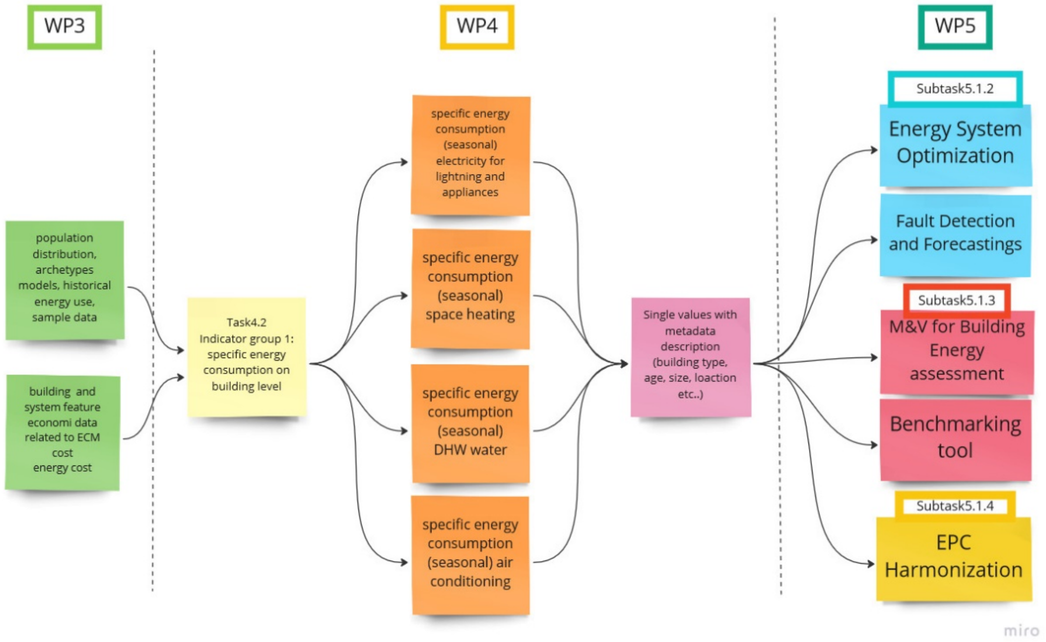

WP3
WP4
WP5
population distribution, archetypes models, historical energy use, sample data
building and system feature economi data related to ECM cost energy cost
Task4.2 Indicator group 1: specific energy consumption on building level
specific energy consumption (seasonal) electricity for lightning and appliances
specific energy consumption (seasonal) space heating
specific energy consumption (seasonal) DHW water
specific energy consumption (seasonal) air conditioning
Single values with metadata description (building type, age, size, loaction etc..)
Subtask5.1.2
Energy System Optimization
Fault Detection and Forecastings
Subtask5.1.3
M&V for Building Energy assessment
Benchmarking tool
Subtask5.1.4
EPC Harmonization
miro

Figure 24: Linkage between Task 4.2, WP3 and WP5.

Figure 30 shows the interlinkage of Task 4.3[80] with WP3 and WP5. The goal of Task 4.3 is to generate synthetic load profiles for electricity and heating (if available). To do so the input from WP3 to Task 4.3 is focused on such profiles. We expect that most profiles which will be processed on the MODERATE platform when finished, will not contain sensible information. However, one key aspect of this Task is to ensure that datasets with sensible information will be processed in a way that the sensible information will not be included in the dataset anymore and at the same time the value of the data is

[79] Create synthetic data of specific energy consumption values on building level.

[80] Create synthetic data of smart meter load profiles on building level.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

not compromised significantly. For training and developing a model which will generate the synthetic data, we expect also input data with some meta data information such as the type of building, number of residents, information on appliances or EPC data. The synthetic data comprises of either yearly hourly arrays typical daily, weekly or seasonal arrays. These profiles can then be further used by WP5 in subtask 5.1.2 in Fault Detection and Forecasting, Energy System Optimization, in subtask 5.1.3 in M&V for Building Energy Assessment and in subtask 5.1.4 EPC Harmonization.

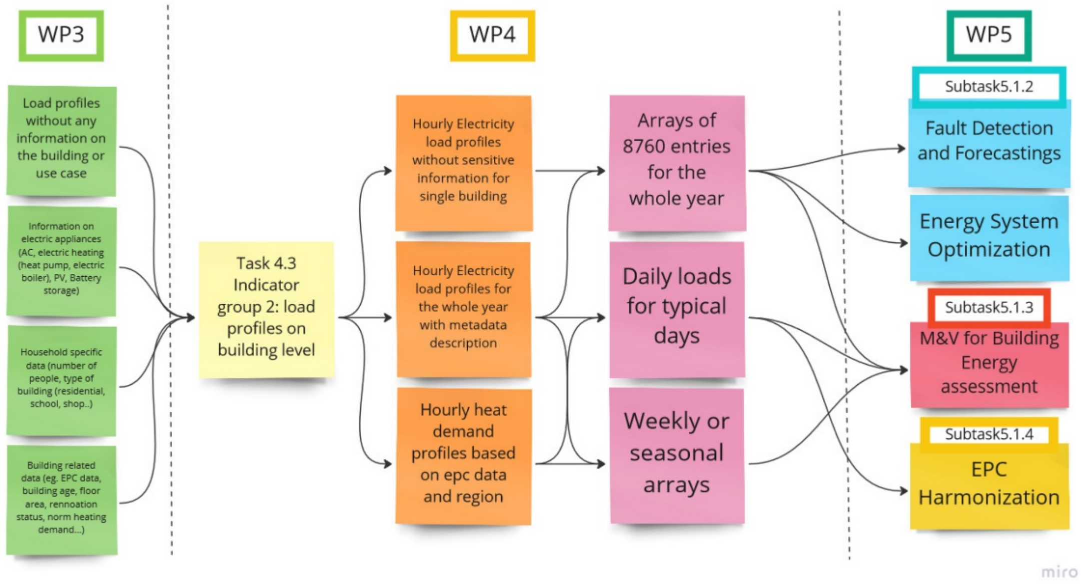

WP3
Load profiles without any information on the building or use case
Information on electric appliances (AC, electric heating (heat pump, electric boiler), PV, Battery storage)
Household specific data (number of people, type of building (residential, school, shop..)
Building related data (eg. EPC data, building age, floor area, rennoation status, norm heating demand...)
Task 4.3 Indicator group 2: load profiles on building level
WP4
Hourly Electricity load profiles without sensitive information for single building
Hourly Electricity load profiles for the whole year with metadata description
Hourly heat demand profiles based on epc data and region
Arrays of 8760 entries for the whole year
Daily loads for typical days
Weekly or seasonal arrays
WP5
Subtask5.1.2
Fault Detection and Forecastings
Energy System Optimization
Subtask5.1.3
M&V for Building Energy assessment
Subtask5.1.4
EPC Harmonization
miro

Figure 25: Linkage between Task 4.3, WP3 and WP5

Figure 31 represents the interlinkage of Task 4.4[81] with WP3 and WP5. Aggregation of data on single building level to regional and disaggregation of data on regional to single building level is the main goal of Task 4.4. In addition, a special focus lies in allocating Heating, Ventilation, and Air Conditioning (HVAC) systems in a given area. Therefore, Task 4.4 relies on both, bottom up data which can consist of EPC data for example and top down data which mostly is provided in the form of statistical data. Various methodologies are used in Task 4.4 to enrich a given database on building characteristics, create data driven archetypes and consumption profiles and break down the energy usage of given buildings (Chapter 3). The results are mainly relevant for Subtask 5.1.2 ECM Application and Subtask 5.1.3 EPC Harmonization and the geo clustering tool.

[81] Aims at aggregating and disagregating building information as well as HVAC identification.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

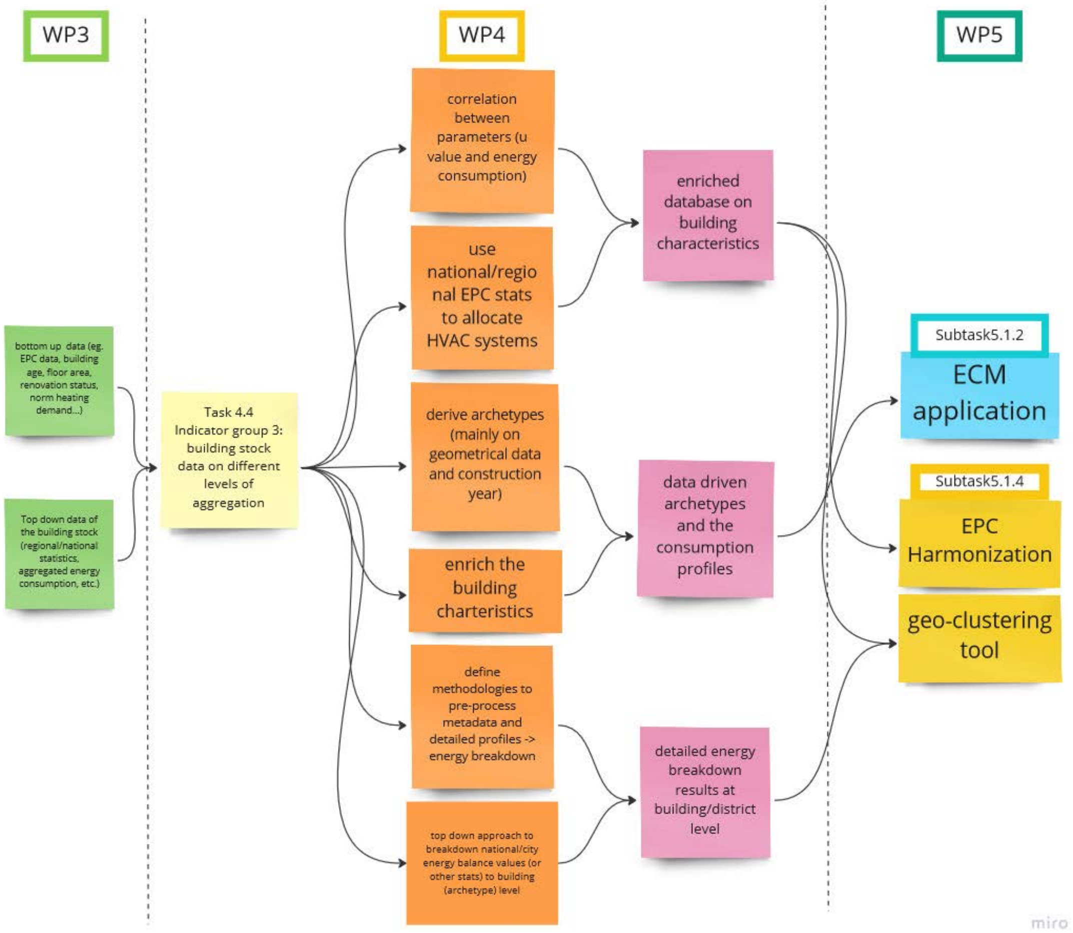

WP3
WP4
WP5
bottom up data (eg. EPC data, building age, floor area, renovation status, norm heating demand...)
Top down data of the building stock (regional/national statistics, aggregated energy consumption, etc.)
Task 4.4 Indicator group 3: building stock data on different levels of aggregation
correlation between parameters (u value and energy consumption)
use national/regional EPC stats to allocate HVAC systems
derive archetypes (mainly on geometrical data and construction year)
enrich the building charteristics
define methodologies to pre-process metadata and detailed profiles -> energy breakdown
top down approach to breakdown national/city energy balance values (or other stats) to building (archetype) level
enriched database on building characteristics
data driven archetypes and the consumption profiles
detailed energy breakdown results at building/district level
Subtask5.1.2
ECM application
Subtask5.1.4
EPC Harmonization
geo-clustering tool
miro

Figure 26: Linkage between Task 4.4, WP3 and WP5.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101069834. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

MODERATE

OUR TEAM

eurac research

Università Ca'Foscari Venezia

Politecnico di Torino

CTIC centro tecnológico

vito

ethink Energy Research

TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna | Austria

Köhler & Meinzer

FONDAZIONE links PASSION FOR INNOVATION

VEOLIA

enercoop GRUPO

ubik Geospatial Solutions

WÜRTH

IVE INSTITUTO VALENCIANO de la EDIFICACIÓN

REHVA Federation of European Heating, Ventilation and Air Conditioning Associations

INSOMNIA

UCLouvain Louvain research institute for Landscape, Architecture, Built environment

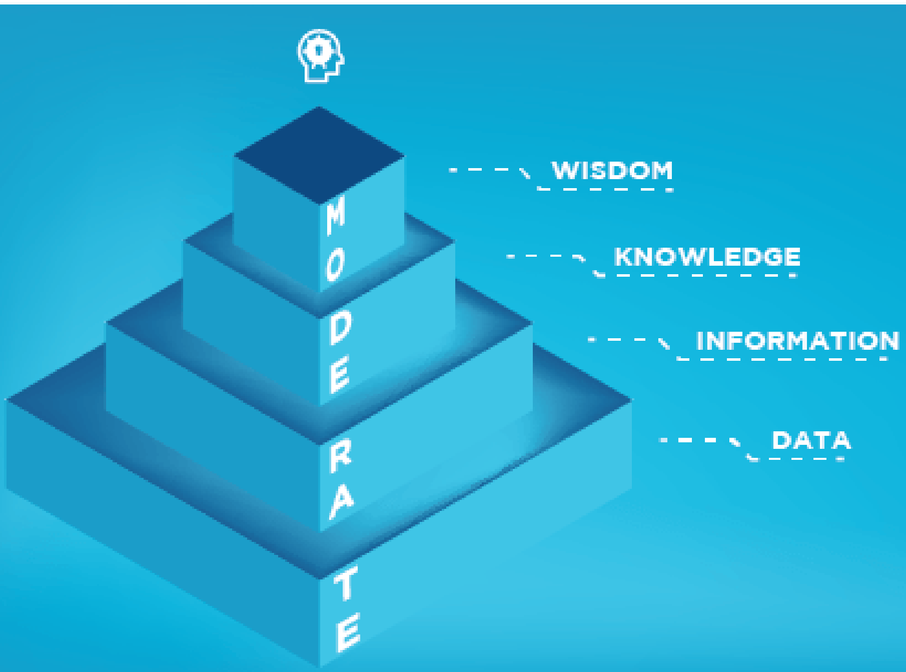
synavision Perfect Building Performance

See you online!

moderate-project.eu

@MODERATE_HE

MODERATE



MODERATE