European Commission

Data-driven method for unsupervised electricity consumption characterisation at district level and beyond

ELISE Energy and Location Applications
Final Report

Mor, G. (CIMNE)

Editors: Martirano, G., Pignatelli F., (JRC)

2021

ELISE

Enabling digital government through geospatial & location intelligence

Joint Research Centre

Contact information
Name: Francesco
Address: Pignatelli
Email: Francesco.PIGNATELLI@ec.europa.eu
Tel.: +39. 033278-6319

EU Science Hub
https://ec.europa.eu/jrc

JRC124888

PDF ISBN 978-92-76-40553-5 doi:10.2760/362074

Luxembourg: Publications Office of the European Union, 2021

How to cite this report: Mor, G., Data-driven method for unsupervised electricity consumption characterisation at district level and beyond — ELISE Energy and Location Applications — Final Report, Martirano, G., Pignatelli F., eds., Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40553-5, doi:10.2760/362074, JRC124888.

Contents

Foreword

This document is the final report of the activities executed in the context of the Contract CT-EX2017D306558-102.

Acknowledgements

Several experts conducted analyses, shared opinions and participated during the design, implementation and validation phases of the presented methodology and, therefore, provided valuable input to this document:

Jordi Cipriano
Eloi Gabaldón
Gerard Laguna
Florencia Lazzari
Stoyan Danov
Benedetto Grillone
Francesc Contreras
Jordi Carbonell
Chiara Lodi
Hans Bloem
Edgar Alexis Martínez
Josep Mayos

Special thanks to Lorena Hernández Quirós (JRC B.6) for her thorough review.

Authors

Gerard Mor is a data scientist from CIMNE - BEE Group, with great experience in building, energy and environment dynamic modelling for forecasting and characterisation purposes.

Abstract

Enhancing energy efficiency has become a priority for the European Union. Several policies and initiatives aim to improve the energy performance of buildings and collect data of sufficient quality on the effect of energy efficiency policies on building stocks across Europe. Knowledge about the characteristics of the building stock and the usage of these buildings' occupants is essential for defining and assessing strategies for energy savings.

Nowadays, dynamic measured data from the Advanced Metering Infrastructure (AMI), especially in electricity consumption, combined with location-based data, like weather, cadastre, social or economic conditions, should be available for a significant part of the building stocks in Europe. Combinedly, this enormous set of data contains the characteristics of how buildings and their occupants consume energy.

In this document, a bottom-up electricity characterisation methodology of the building stock at the local level is presented. It is based on the statistical analysis of aggregated energy consumption data, weather data, cadastre, and socioeconomic information. For validation purposes, the characterisation of the electricity consumption over Lleida (Spain) province is performed. The geographical aggregation level considered is the postal code (more detailed than LAU level 2, formerly NUTS level 5), due to it is the highest resolution available through the Spanish Distribution System Operators (DSOs) data portal. Besides, a web application to visualise the results of the characterisation has also been developed. The major novelty is the use of high-frequency consumption data from most consumers in each analysis area without considering any Building Energy Simulation (BES) model that considers performance or energy use assumptions. For this purpose, a data-driven technique is used to disaggregate consumption due to multiple components (heating, cooling, holiday and baseload). In addition, multiple Key Performance Indicators (KPIs) are derived from these components to obtain the characterisation results. The potential reuse of this methodology allows for a better understanding of the drivers of electricity use, with multiple applications for the public and private sectors.

This study has been executed in the frame of the Energy & Location Applications of the ELISE (European Location Interoperability Solutions for e-Government) action of the ISA^2 (Interoperability solutions for public administrations, businesses and citizens) Programme.

Executive Summary

The Digital Economy Unit of the European Commission's Joint Research Centre (JRC), in cooperation with other services of the European Commission, is coordinating the "European Location Interoperability Solutions for e-Government (ELISE)", Action 10 of the ISA^2 (Interoperability Solutions for Public Administrations, Business and Citizens) Programme.

The ISA^2 Programme supports long-standing efforts to create a European Union free from electronic barriers at national borders. It facilitates interaction between European public administrations, businesses, and citizens to enable interoperable cross-border and cross-sector public services and ensure the availability of common framework and solutions.

The ELISE action is a package of guidance and solutions facilitating efficient and effective electronic cross-border or cross-sector interactions between European public administrations, citizens and businesses, in the domain of location information and services.

The ELISE Energy & Location Applications consist of a series of use cases aiming to show how location data can support different stakeholders involved in the energy policy cycle at different geographical scales, ranging from local to European-wide level.

This report focuses on the use case named "Scale-up methodologies", aimed to identify, apply and test in a real case study the most appropriate methodologies and technologies to scale-up at district, city, regional and national levels energy efficiency assessments made at building level based on dynamic measures of energy consumption.

Energy characterisation of existing buildings at multiple geographical levels (district, city, and region) can be used to understand trends in energy use, to correlate the energy consumption to characteristics of the territory and to identify specific locations where there are buildings with poor energy performance. But it is often difficult to obtain this characterisation, which can be tackled from different points of view, with widely varying levels of accuracy and associated costs. In this use case, a bottom-up approach, based on a data-driven methodology, addresses the problems related to this characterisation. In essence, the methodology consists in analysing the energy consumption and the weather data at district level to obtain KPI's and correlate these indicators with building stock and socioeconomical characteristics, to obtain the energy trends and characteristics of each district. This methodology was implemented and validated within the Spanish province of Lleida and the main data sources used are: electricity consumption datasets aggregated at postal code level, INSPIRE harmonized datasets of buildings, weather data from online services and socioeconomical datasets from the National Statistics Institute. The main goal was to provide a characterisation of the electricity consumption, either in the residential or in the public sector, at different geographical levels (Spanish census tract and postal code data are used for validation purposes).

In this report, a data-driven methodology addressing the problems related to this characterisation is presented.

The benefits expected from the use of this methodology are to support the public (e.g. Regional Energy Agencies) and the private sector (e.g. Distribution System Operators (DSOs), utilities, ESCO's, and companies working in the building renovation sector) to understand energy efficiency and energy use trends and their relation with building stock characteristics and socioeconomic factors. Indeed, this methodology should allow correlating the energy usage and the territory massively, providing a better understanding and interpretability of a region, enabling to detect energy efficiency-related faults and business opportunities. Additionally, a second benefit of this methodology is that it could be applied over large areas (a whole province, or an entire region or even a country), once the required public data would be available, which seems to be possible for several EU countries thanks to the INSPIRE data harmonisation and the steady growth of open data platforms. Compared to other methodologies: no simulations, calibration procedures, user surveys, or significant human interaction are required. Thus, it can be deployed with less effort in terms of both computational and human resources.

1 Introduction

Characterising the energy demand at the local level is essential to understand the dynamics that support the transition to renewable and distributed generation areas. A recent study [1] has shown the necessity to explore energy efficiency solutions for buildings at the local aggregated level (e.g. district, neighbourhood, city, region). The implementation of local Energy Conservation Measures (ECM) and the increase of in-situ renewable generation in buildings are key factors to satisfy energy security and limit global warming in future. This local geographical level is large enough to infer a priori unknown patterns of energy consumption and addresses concrete solutions or, at least, help decision-making teams in energy planning scenarios. Additionally, this is the geographical scale where most urban transformations in Europe occur and where the newest instruments for financing energy efficiency strategies in the building sector exist.

In literature, the energy characterisation based on modelling in groups of buildings is called building stock modelling. Three major typologies of buildings exist residential, industrial and services. Each of them corresponds to its building archetypes, uses and occupancy patterns. Mainly, two approaches for building stock modelling can be identified: top-down and bottom-up methods. Lagevin et al. [2] provide a large and updated literature review to extend this initial classification, considering three major developments during the last ten years: big data, increased computing power, and new modelling techniques.

A bottom-up approach begins with a detailed representation of a system's constituent part that may be aggregated up to the whole-system level. In this case, it will look at individual buildings through either the use of the distinct archetypes, or the use of each building of the stock individually, or, at least, a sample of buildings to aggregate the total stock.

By contrast, top-down approaches begin with an aggregated view of the system that could be disaggregated into subsequent sub-systems. In the case of energy characterisation of groups of buildings, they analyse the building stock as a large sink with inputs and outputs following historical trends and therefore keeps the building stock as a black box (Langevin et al. 2020). Within the two main approaches, i.e. top-down and bottom-up, subdivisions can be made.

A data-driven methodology for urban electricity demand modelling was tested for Dutch municipalities, where a combination of multiple datasets (reference electricity demand profiles, local customers composition data, and aggregated local annual demand data) was used to train a regression model for local electricity demand prediction with an application for the Local Energy Transition [3].

To sum up, energy characterisation of existing buildings at multiple geographical levels (district, city, region) can be used to understand energy use trends, correlate the energy consumption to characteristics of the territory, and identify specific locations where there are buildings with poor energy performance. Nonetheless, it is often difficult to obtain this characterisation, which can be tackled from different points of view, with widely varying levels of accuracy and associated costs.

Typically, in the case of bottom-up approaches, the characterisation of the energy performance of a given region is performed using building energy simulation (BES) models. In these cases, a calibration of the simulated data against real monthly or annual consumption should be considered, which aims to avoid performance gaps between reality and simulations. Nonetheless, this type of calibration procedure usually ignores the changes in the behaviour of the users over time, and in some cases, the dynamics between the real consumption and the climate conditions. Moreover, in several methodologies, a subset of representative buildings should be considered to depict the archetype of a certain region. Hence, large biases against reality could be done in the sampling or the calibration procedures. All these factors result in high inaccuracies in the estimates of energy performance.

A statistical methodology that uses the actual data to estimate the KPIs of the energy consumption without additional calibration procedures, sampling of buildings, or input data from customers, would be interesting, especially in terms of scalability. The expected benefits are to support the public (e.g. Regional Energy Agencies) and the private sector (e.g. DSO's, utilities, ESCO's, and companies working in the building renovation sector) to understand energy efficiency and energy use trends and their relation with building stock characteristics and socioeconomic factors. Indeed, this methodology should allow correlating the energy usage and the territory massively, providing a better understanding and interpretability of a region, enabling to detect energy efficiency-related faults and business opportunities. Additionally, a second benefit of this methodology is that it could be applied over large areas (a whole province, or an entire region or even a country), once the required public data would be available, which seems to be possible for several EU countries thanks to the INSPIRE data harmonisation and the steady growth of open data platforms. Compared to other methodologies: no

simulations, calibration procedures, user surveys, or significant human interaction are required. Thus, it can be deployed with less effort in terms of both computational and human resources.

In this report, a data-driven methodology is presented to address the problems related to this characterisation. In essence, it consists of the correlation of the energy consumption, weather data, and the building stock and socioeconomic characteristics to obtain the normalised energy trends and KPIs that describe the consumption of each district.

Nevertheless, some issues still exist nowadays regarding the availability of energy consumption datasets at the needed aggregation levels, both in terms of geographical resolution and time frequency. Therefore, taking into account the higher implementation in certain EU countries of the Advanced Metering Infrastructure (AMI) in electricity consumption, it is much more feasible to obtain detailed sets for this type of energy than for the rest (gas, biomass, oil). In summary, and as a first validation of the data-driven characterisation methodology presented in this report, electricity consumption has been considered the only energy resource to be characterised due to the problems in obtaining detailed data for the other main energy resources used in EU buildings. Thus, the requirements are the aggregated hourly electricity consumption by type of use (residential, industrial or services) and the related building stock, meteorological data and socioeconomic conditions of each geographical area in analysis.

The methodology is validated within the Spanish province of Lleida, and the main data sources used are hourly electricity consumption datasets aggregated at postal code level from the Datadis platform, INSPIRE harmonised datasets of the Spanish cadastre, weather data from an online service and socioeconomic datasets from the Spanish National Statistics Institute. The main goal is to provide a geographically aggregated characterisation in terms of building performance and usage trends of the electricity consumption, both for the residential and public/tertiary buildings.

ELISE Webinar: Data-driven methodology for electricity characterisation of districts

The methodology and its application described in this report have been also presented in the ELISE Webinar: Data-driven methodology for electricity characterisation of districts, held on 15.07.2021. Further details and the webinar recording are available in the dedicated JoinUp page.

1.1 Preliminary approaches of the methodology

During the development, implementation, and validation of the characterisation methodology, several approaches were discussed from December 2019 to February 2021, and new datasets become publicly available during that period.

In an initial phase, aggregated electricity consumption from Datadis platform was not available. Hence, the characterisation at a certain geographical level was intended to be an aggregation of several individual characterisations of a representative number of consumers, sufficient to represent all customers' general distribution at that level. The focus groups were the residential sector and the public buildings sector. In the first case, CIMNE could provide a great amount of residential consumption datasets for several postal codes located around Barcelona and Girona (Spain). For the second one, additionally to the CIMNE public buildings datasets gathered during several European projects in partnership with the Government of Catalonia, the consumption data of public buildings available through open data sites could also be considered. Although the procedure was different from the one finally implemented and validated, the original goal remains unchanged from the beginning of the project: to exploit the usage of metering data and INSPIRE harmonised datasets to provide a characterisation of energy consumption in the residential and public buildings sector to support government and private entities regarding energy efficiency and energy use trends.

As mentioned in the introduction, an attempt has been made to include energy resources such as gas, biomass and oil in the analysis. Then, the interpretation of the characterisation could be understood as the performance of the buildings and their occupants against the total energy consumption produced in the buildings, regardless of the rate of implantation of the different energy resources in the building equipment (heating boilers, chillers, cooking equipment, domestic hot water). Nonetheless, the actual availability of big datasets containing high-frequency gas, biomass or oil consumption is extremely low, especially for the residential sector. This point is very significant in Spain, where the validation was conducted, and only electricity consumption data is really available for a considerable number of customers. In the mid and long term, this fact should evolve positively to implement energy data-driven characterisation techniques due to the pronounced tendency to electrify all-kind of building systems and the strong implementation of advanced meters for gas consumption.

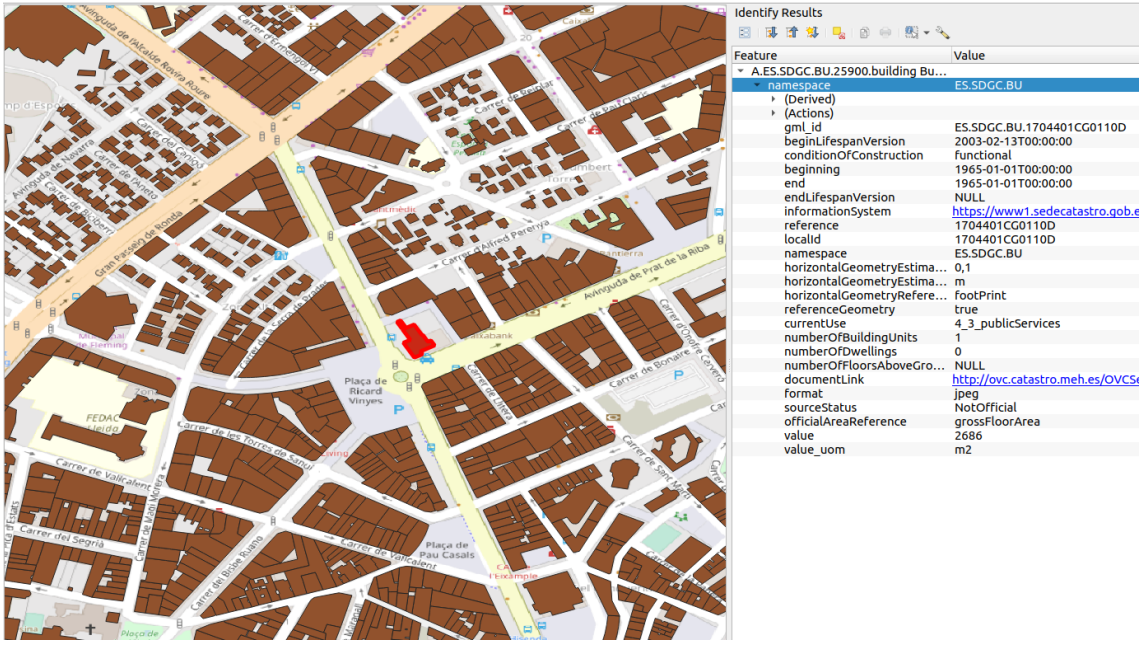
2 Description of the characterisation methodology

The data-driven methodology presented in this report aims at characterising the electricity consumption of large areas starting from the district level (postal code, in the case of Spain) to upper levels. In essence, it obtains data from multiple sources and aggregates this information to the least common geographical level to merge the information and provide statistical models used to estimate the drivers of electricity consumption related to location. The outcomes are shared through a web dashboard that depicts the detailed electricity consumption characterisation for each postal code and an interactive map to benchmark the complete set of indicators between all postal codes. As can be seen in section 3, the validation of this methodology is tested over the Spanish province of Lleida (>12500 km^2) [4].

2.1 Data collection and refinement

2.1.1 Cadastre data

Figure 1. Spanish Cadastre layer visualised in QGIS



Source: CIMNE - BEE Group own elaboration and Open Street Maps

In the context of this project, building information is gathered from the national cadastral datasets. The data format used by these entities across EU countries is harmonised using the INSPIRE Buildings theme [5]. In the case of Spain, the massive downloadable public information of cadastral datasets is available through ATOM files [6], where the GML files regarding "buildings" and "building parts" can be obtained for all the municipalities (see **Figure 1**).

Basically, those files contain a set of georeferenced information for each building and, depending on the type of information described; each variable could be grouped in:

1. Geometry information contains knowledge about 2D geometries of the building parts, gross floor area, number of floors above and below ground.
2. Typology information, which contains variables, such as the major current use, the total number of dwellings and building units.
3. Construction information, which contains the actual conditions of the building and the year of construction.

Even if the amount of information is really extensive, it has to be considered that multiple drawbacks exist when using cadastral data gathered through ATOM files. In the case of the variables belonging to groups 2 and

3 described before, it should be considered that many data inaccuracies can exist compared to real conditions. Some of the encountered issues are described in the following:

- Problems when dealing with buildings with several main uses (services + residential, or industrial + services).
- Non-realistic dwelling areas based on the gross floor area due to the influence of large parking and/or community areas.
- Some of the building information not available for all the regions (Buildings located in the countryside vs. those located in cities). For instance, in certain rural areas of the Lleida province, there's up to 30% of buildings without current use information.

In order to avoid unrealistic estimations when the cadastre information is aggregated to district geographical levels and beyond, some filters where considered.

2.1.2 Socioeconomic data

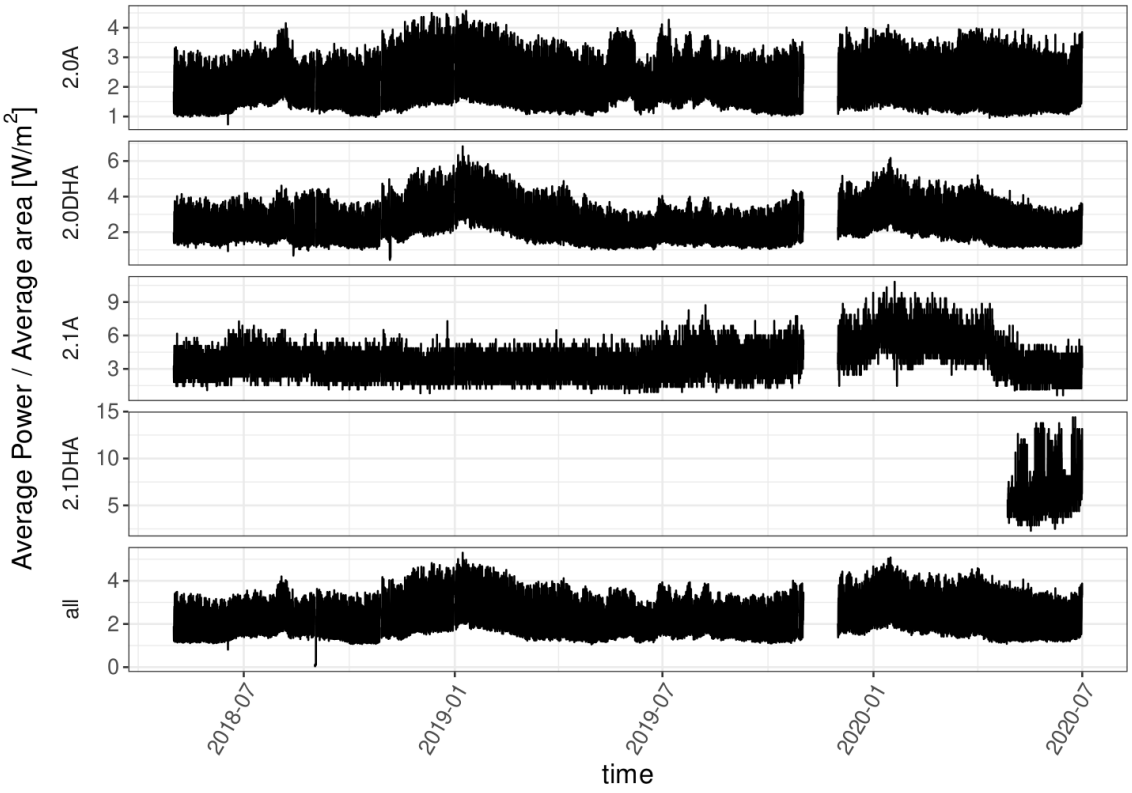
The economic status and the demographics indicators considered in this methodology are gathered through national statistics institutes. In the case of Spain, this data can be obtained from an experimental project of the Spanish Statistical Office (INE), named "Household income distribution map" [7]. This project proposes constructing statistical indicators of the level and distribution of household income at the municipal and census tract geographical levels from the link between INE's demographics information and the tax data from the National and the Autonomous Treasuries.

Some of the indicators obtained at the census tract geographical level are the average incomes per person and household, incomes main sources, incomes quantile 80 and 20 ratio, number of inhabitants, average population age, percentage of people under 18 and over 65, number of people per household, percentage of single households, and Gini index.

2.1.3 Electricity consumption data

In case of bottom-up characterisation methods, the electricity consumption of a representative sample of residential and public buildings should be used to estimate at an individual level. Afterwards, aggregate the individual characterisation indicators to the geographical level of interest. Nonetheless, mainly due to data protection laws, it is unfeasible to obtain public or freely available datasets with significant consumption points over large geographical areas. Therefore, the consumption data used to validate the presented methodology is natively aggregated to the postal code geographical level. In the case of Spain, Datadis web platform [8] supplies the historical hourly electricity consumption aggregated by postal code, economic sector, tariff and DSO for the whole country. Most Spanish DSO's participate in this platform, which provide electricity services to around 28 million consumption points. The aggregated hourly consumption is gathered through the Datadis API, which requires authentication using an FNMT electronic certificate [9] of a legal entity. On average, most of the postal codes contain two years of historical data. The aggregated information for each obtained item through the API (a single day, postal code, economic sector, tariff, and distributor company) is the total consumption, the consumption at each hour of the day and the number of online contracts.

Figure 2. Aggregated electricity consumption of a subset of tariffs in a single postal code



Source: CIMNE - BEE Group own elaboration and Datadis

Figure 2 depicts the average hourly residential electric power per built area and tariff in the postal code related to *Pardinyes* neighbourhood in the city of Lleida. As it can be seen, data contains a significant gap during November 2019, and multiple energy trends and seasonality's exist between different tariffs. Due to this fact, a synthetic tariff is created, named "all", weighting its values using the number of contracts per each of the tariffs. This aggregated tariff improves the representativity of each postal code when the results are visualised over a map.

In Spain, during the period of data represented in this report (2018-2019-mid 2020), the electricity tariffs framework for a contracted power of less than 450kW is the following:

Table 1. Electricity tariffs description in the Spanish market

Access toll name	Time-of-use structures	Time of use	Contracted power range (and voltage level)
2.0	A	No	<10kW (Low voltage)
	DHA	Yes (2 periods)	
	DHS	Yes (3 periods)	

2.1	A	No	>10kW and <15kW (Low voltage)
	DHA	Yes (2 periods)	
	DHS	Yes (3 periods)	
3.0	A	Yes (3 periods)	>15kW (Low voltage)
3.1	A	Yes (3 periods)	<450kW (High voltage)

Source: Red Eléctrica Española (REE)

In general, medium and small-sized residential dwellings and small shops use tariffs 2.0, which allows up to 10 kW of contracted power, and offers three time-of-use structures: A (fixed during the whole day), DHA (Lower energy price during night and morning), and DHS (3 periods, lower energy price during the night, medium during morning and evening, and higher during afternoon). DHS time-of-use structure is normally useful for people who have electric vehicles. In big-sized houses and medium-sized shops or small offices, users tend to contract tariffs 2.1, which allows a contracted power between 10 and 15 kW with the same time-of-use tariffs than the 2.0 access toll.

The 3.0 tariff is usually contracted in very singular residential houses with lots of electric equipment, commercial buildings and medium-sized industries. Finally, the 3.1 tariff and beyond are contracted by big industries, normally located in large industrial areas, as they need high voltage energy distribution for their applications.

2.1.4 Weather data

Outdoor weather conditions are obtained through the Darksky API service for the whole area in analysis. In essence, the historical weather data for the same period is downloaded for each one of the postal codes considered. The most important variables in our analysis are the outdoor temperature and wind speed. With these variables, a data-driven model is used to infer, if it exists, the weather dependence of the electricity consumption during the heating and cooling periods.

2.2 Geographical levels

Data used in the framework of this energy characterisation is related to multiple geographical levels. In this section, each of the available geographical levels is described. Moreover, in the data integration section, it is described how all data sets are normalised to the same level, which is a necessary step to analyse the datasets.

2.2.1 Building level

Data referenced to this level contains the exact location where the building is physically placed. Cadastral data is an example of a dataset with this geographical level. Apart from cadastral information, and mainly due to privacy issues, there are few other open datasets available at this level. Nonetheless, it is worth mentioning that this geographical level would be the most interesting one due to its flexibility in terms of aggregation. For instance, characterisation results could be easily aggregated by streets, blocks of buildings, neighbourhoods or custom aggregations which could provide differences within the census tract or postal code levels.

2.2.2 Postal code level

The postal code is a code that is assigned to different areas or places in a country. Initially, it was a code to facilitate and mechanise the delivery of mail. It usually consists of a series of digits, although in some countries, it includes letters. In the case of Spain, it is composed of the province code (two first digits) and then three more digits which represent each different postal code. The institution that defines them is the "Sociedad Estatal

Correos y Telégrafos, S.A.". Many other companies, or even the government, widely use this geographical level to refer their data to its location; it strikes a good balance between anonymity, simplicity and detail.

The shape of each postal code is obtained from KML files stored in codigospostales.com [10].

2.2.3 Census tract level

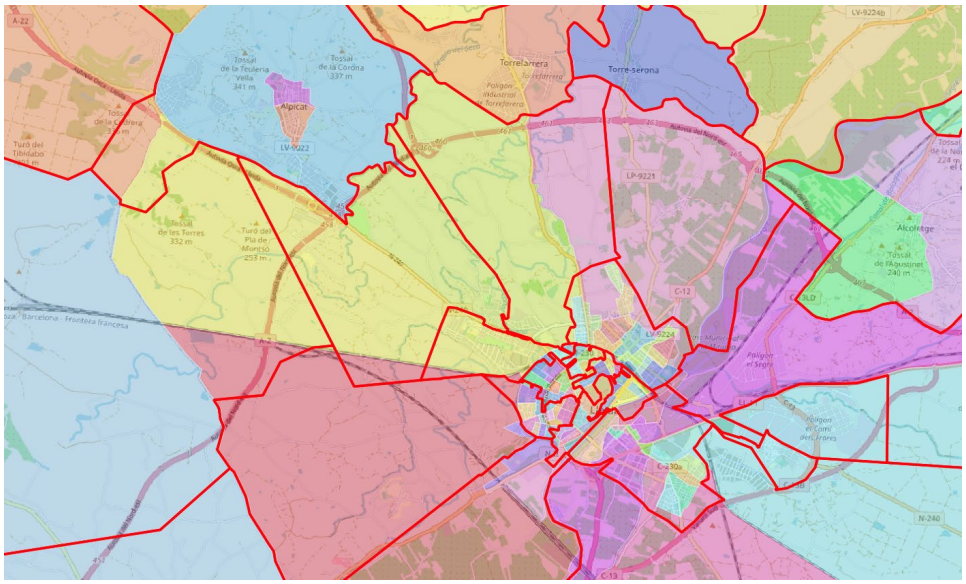
Census tracts are the lowest level units for disseminating statistical information (e.g. from censuses) and are also used to organise electoral processes. Being operational in nature, they are always defined by more or less fixed sizes: the number of statistical surveys that an interviewer agent can distribute and collect for population counting purposes in the time of one or two months, or the number of people who can vote in a ballot box without crowding on an election day.

The multiple purposes of the sections require special attention to be paid to their demarcation and size. Therefore, they are defined by easily identifiable boundaries (rivers, streets, etc.) and have a size of between 1,000 and 2,500 inhabitants, unless the municipality concerned has a smaller population. The size is given by the Electoral Regime Law, which assigns a minimum and maximum population measured in number of electors. In addition, it is recommended that the size of a section should not exceed 2,500 inhabitants to facilitate the operation of the pollsters.

This law determines that it is up to the provincial delegations of the Electoral Census Office to establish the number and limits of the census sections. Normally, the National Statistics Offices use the census tracts as the lower level of their statistical analysis to avoid privacy issues when dealing with private data that must be anonymised.

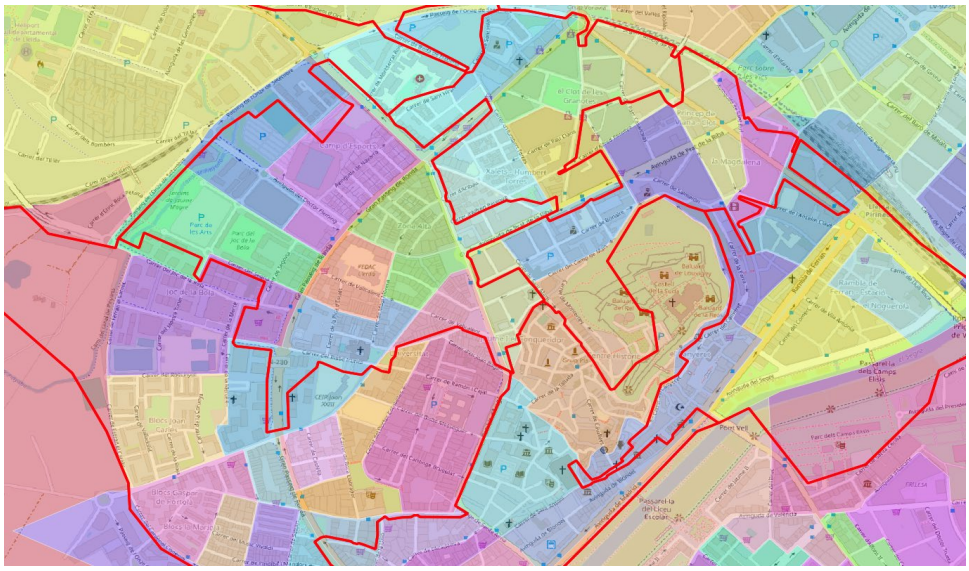
The most updated shape files of the census tract in Spain are obtained from the National Statistical Office [11].

Figure 3. Representation of the census tract (colour-filled) and postal code (red-framed) geographical levels in a mixed urban-rural region



Sources: CIMNE - BEE Group own elaboration and OpenStreetMap

Figure 4. Representation of the census tract (colour-filled) and postal code (red-framed) geographical levels in an urban region



Sources: CIMNE - BEE Group own elaboration and OpenStreetMap

In **Figure 3** and **Figure 4**, the differences between the geographical levels are depicted. The census tract polygons are painted in multiple colours, whereas postal codes are the red-outlined areas, and the buildings are represented in a darker tone between the streets/roads. The census tract level offers much more detail for urban areas compared to the postal code one. The number of blocks of buildings inside a certain census tract is much lower than in the postal code level. However, for rural areas, the representativity of both levels is very similar, as they usually represent areas of similar size.

2.3 Data-driven electricity characterisation at postal code geographical level

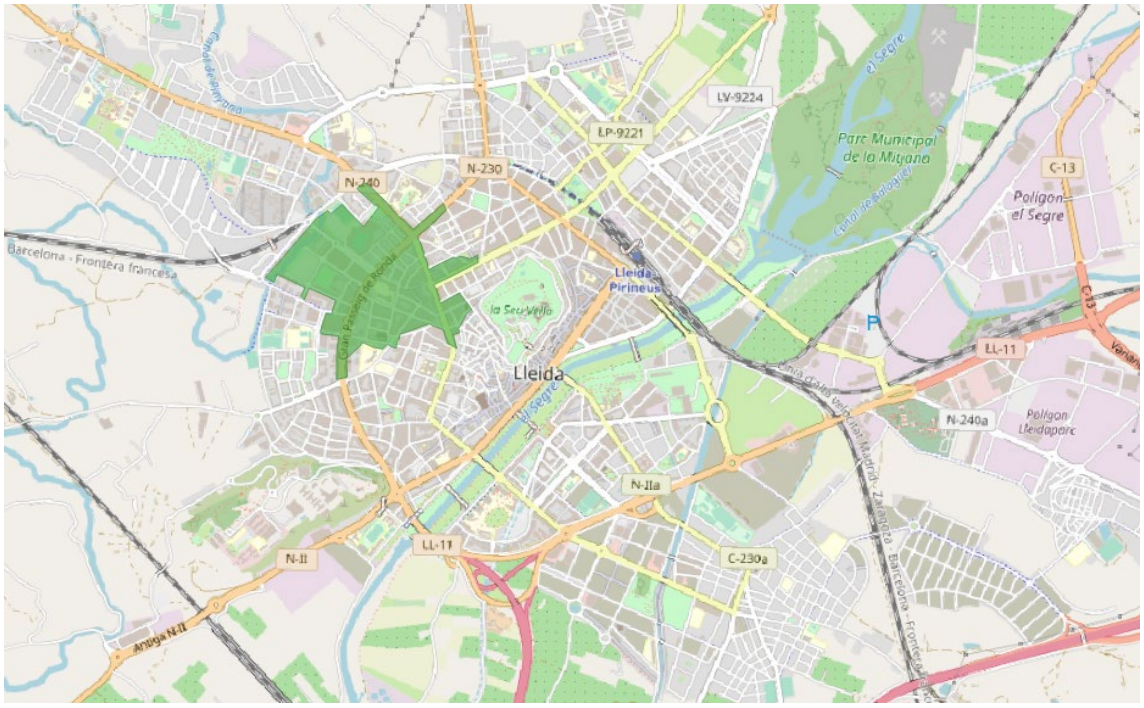
The characterisation methodology consists of the following major steps:

- Data cleaning of the electricity consumption data.
- Clustering the daily load curves to infer the most representative usage patterns.
- Estimate a data-driven electricity consumption model using the calendar, clustering results, and weather conditions as exogenous variables.
- Disaggregate the raw electricity consumption in baseload, heating and cooling components using the model estimated in the previous step.

2.3.1 Case study

To increase the understandability of the following subsections, one postal code is selected to depict intermediate and final results of the characterisation. The selected one is the 25006, which is related to the *Zona Alta* neighbourhood in Lleida (see **Figure 5**). It's known as one of the most well-being districts in Lleida, at least compared to the ones near the city centre. Some of its socioeconomic characteristics are an annual incomes average of 36,498€ per household, an incomes quantile 80-20 ratio of 3.23 (one of the highest of the province), an average population age of 47.42 years old, with 26.95% of people older than 65 and 13.59% under 18.

Figure 5. Postal code selected for the case study



Sources: CIMNE - BEE Group own elaboration and OpenStreetMap

2.3.2 Data cleaning

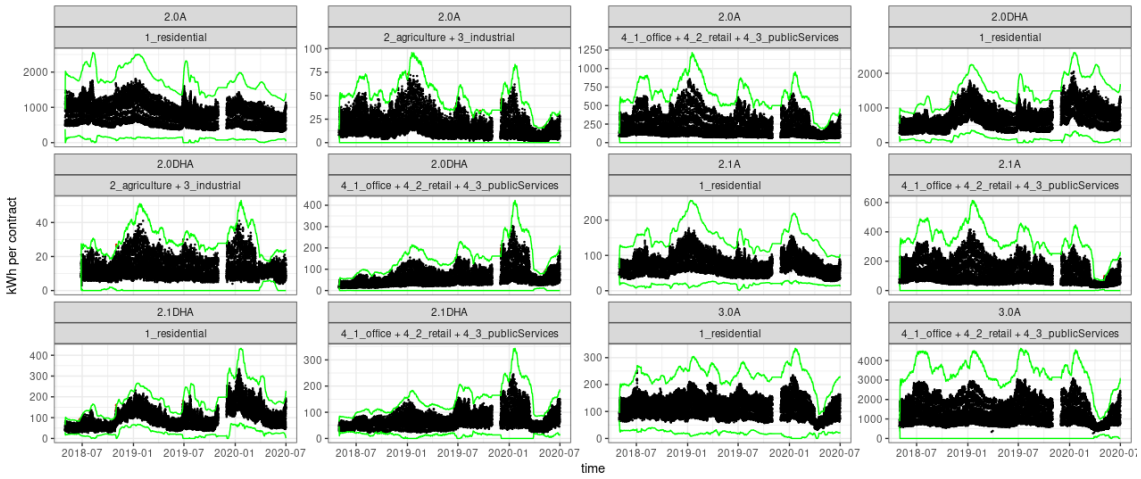
Raw data from electricity distribution companies usually contains several gaps or inaccuracy problems due to metering or communication issues. In this case, even considering the use of aggregated consumption data at a postal code level, which alleviates the influence of poorly measured data at some particular sites, some problems were detected during the initial quality checks. Hence, it became mandatory the implementation of a data cleaning process before the analytics and modelling steps.

One of the following conditions must be accomplished to consider a certain consumption as an outlier:

1. Hourly consumption equal to 0. In the case of electricity consumption aggregated to the postal code level, it is certainly impossible to have a zero consumption considering the number of contracts available per each postal code.
2. Hourly consumption lower than the maximum feasible contracted power, depending on the tariff restrictions. For instance, the contracted power must be lower than 10 and 15 kW, respectively, for 2.0 and 2.1 tariffs.
3. Hourly consumption is six times higher than the 3rd quartile of all the historical consumptions.
4. Hourly consumption outside the right moving average +- 3 moving standard deviations, considering a window of 15 days.

Figure 6 shows the upper and lower bounds describing the limits for the outlier's detection. The multiple combinations of tariffs and economic sectors within postal code 25006 are shown.

Figure 6. Electricity consumption time series of the case study and the outlier's threshold considered in each series in green



Source: CIMNE - BEE Group own elaboration

2.3.3 Inferring usage patterns

Clustering daily load curves can be used to detect similar usage patterns. The representative groups obtained would be used along the algorithm to increase the reliability of the characterisation due to the consideration of the multiple seasonalitys in the building that could not be related to calendar variables, such as different electricity load profiles during weekdays and weekends, or weather conditions.

Clustering can be achieved using various algorithms, which differ in their way to define the constituents of a cluster and how to find them efficiently. The best-suited clustering algorithm depends on the particular data set and the intended use of the results.

In this study, the achieved outcome of the clustering technique is a set of centroid curves for 24h electrical load, which will define the typical load patterns for each user.

The Z-score is calculated for each hour of the day following the equation taking again y_t as the energy consumption data at time t, but the width N, in this case, is equal to the entire length of the data set (no window is applied).

$$Zscore_t = (y_t - median(y_{Nt}))/\sigma_{median(y_{Nt})}$$

Among the different clustering techniques, distribution-based clustering was chosen because it is the one that most closely resembles the way energy measurement data sets are generated by sampling random objects from a distribution.

Let $Y = Y_1, Y_2, Y_3, \ldots, Y_n$ be a sample of n independent identically distributed observations.

In this case, for example, the observation Y_1 is day 1 with its values in the 24 corresponding variables (each hour of the day is a variable), that is to say, $Y_1 = y_{00:00}, y_{01:00}, \ldots, y_{23:00}$. (The graphical interpretation would be a point in a 24th-dimensional space.) The distribution of every observation is specified by a probability density function through a finite mixture model of G components, which takes the following form

$$f(x_i; \Psi) = \sum_{k=1}^{G} \pi_k N(\mu_k, \Sigma_k)$$

where $\psi = \{ \pi_{1,\ldots}, \pi_{G-1}, \mu_{1,\ldots}, \mu_{k,\ldots}, \mu_G, \Sigma_{1,\ldots}, \Sigma_G \}$ are the parameters of the mixture model, $N_k(x_i;\mu_k, \Sigma_k)$ is the k^{th} component Gaussian density for observation x_i with parameter vector (μ_k, Σ_k), and $(\mu_{1,\ldots,}\mu_{G-1})$ are the mixing weights or probabilities (such that $\pi_k > 0, \Sigma\pi_k = 1$, and G is the number of mixture components (in the model-based approach to clustering each component is associated with a group or cluster). Assuming that G is fixed, the mixture model parameters ψ are usually unknown and should be estimated.

In the case described above, it is assumed that all component densities arise from the same parametric distribution family: the Gaussian. Thus, clusters are ellipsoidal, centred at the mean vector μ_k, with geometric features, such as volume, shape and orientation, determined by the covariance matrix Σ_k.

The mixture of multi-dimensional Gaussian probability distributions that best fit the input dataset is estimated via the expectation-maximisation algorithm for maximum likelihood estimation. The covariance (Σk) structures for parameter estimation of Gaussian mixture models are the following:

- Spherical: variance is equal in all directions (where the "directions" are the 24 variables, one for each hour of the day).
- Diagonal: each direction has a different variance.
- Ellipsoidal: allows covariance terms to orient ellipse in different directions plus constraints regarding shape and volume of the Gaussian density functions.

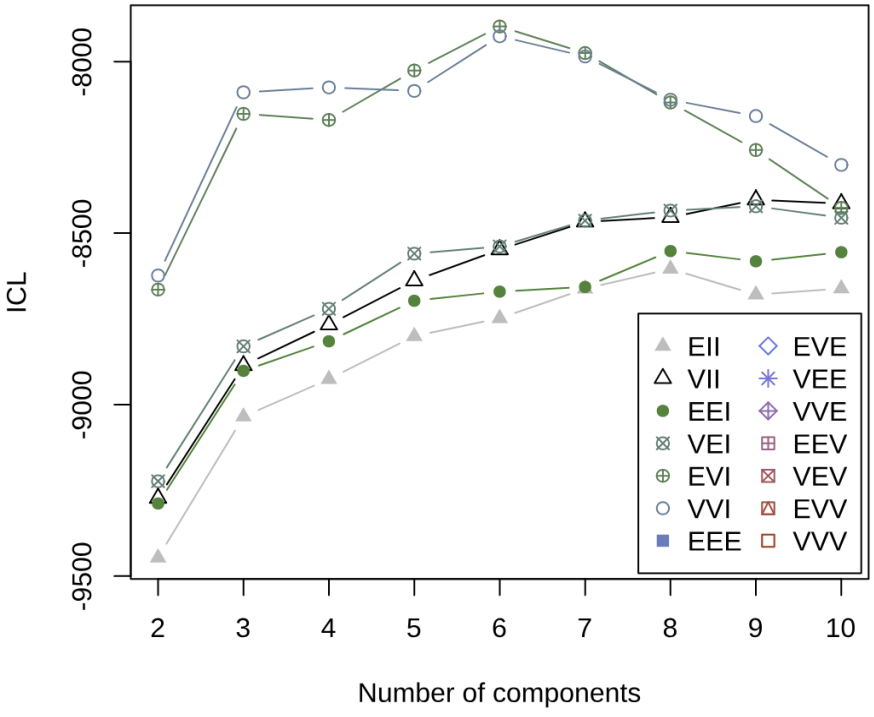
The Gaussian Mixture Model is computed for *G* clusters between 2 and 10.

The optimum total amount of clusters is selected by using the Integrated Completed Likelihood (ICL) criterion and the model fit is done using the Bayesian Information Criterion (BIC).

The key difference between the BIC and ICL is that the latter includes an additional term (the estimated mean entropy) that penalises clustering configurations exhibiting overlapping groups. While BIC allows a very efficient fit for the mixture model, ICL allows a good estimation instead of the number of clusters appearing in the data.

In **Figure 7**, the ICL vs the number of clusters (components) is shown for an arbitrary user; in this case, the model EVI (name for diagonal, equal volume, varying shape) with *G* equal to 6 was selected by the maximisation of the ICL.

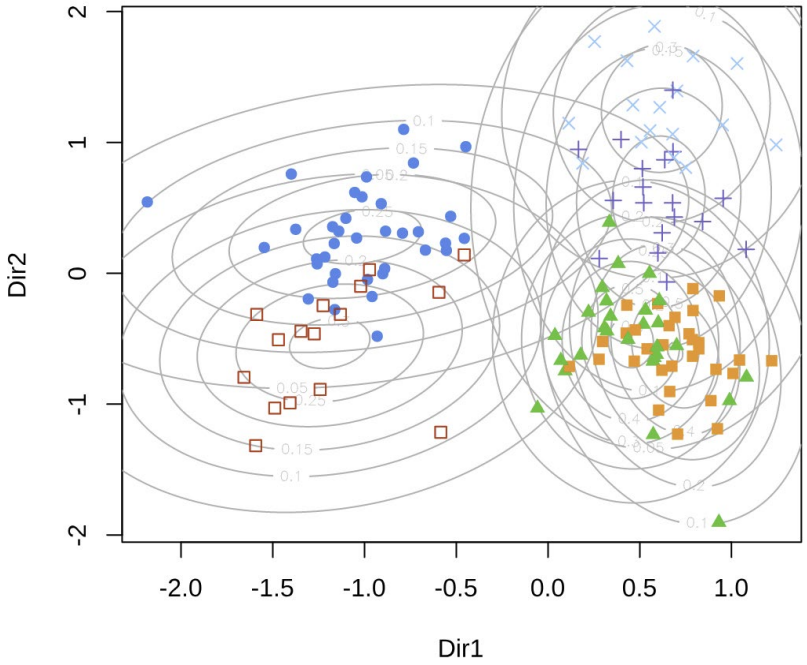
Figure 7. Selection of optimal number of clusters



Source: CIMNE - BEE Group own elaboration

According to reduction dimension theory, the *24G* space may be reduced to one with several directions equal to *min{24, G - 1}*. Using the representation in these directions, the contours of the fitted Gaussian density distributions are shown in **Figure 8**.

Figure 8. Multiple gaussian distributions representing each detected cluster



Source: CIMNE - BEE Group own elaboration

The plot shows the representation in the particular Dir1 x Dir2 space. The ellipses represent the contours of the fitted Gaussian density functions. The data points are painted in different colours and shapes, representing the result of the clustering.

The centroids are the daily load characteristic pattern for each cluster. To obtain the centroids, the following average is calculated over all the days belonging to each cluster:

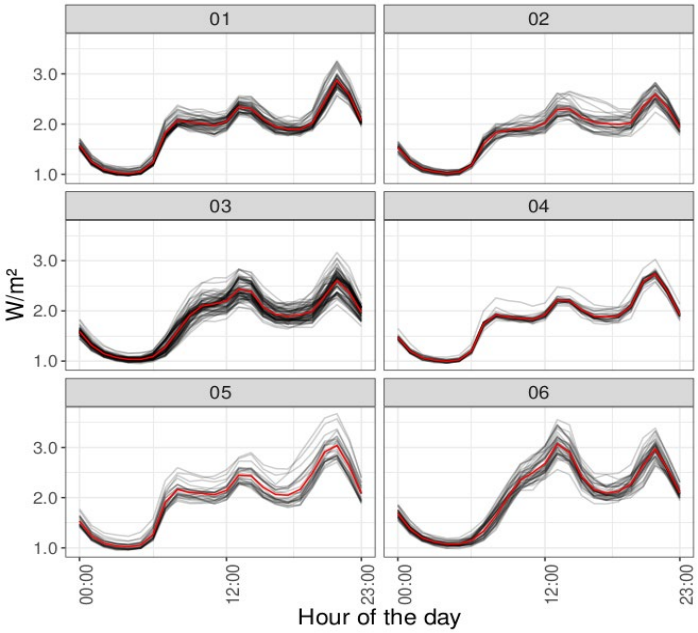
$$c_{s,h} = \frac{1}{n_s} \sum_{i=1}^{n_s} y^i_{s,h}$$

where $c_{s,h}$ is the value for the centroid s at hour h, n_s is the total amount of days belonging to cluster s and $y^i_{s,h}$ is the electricity consumption data (normalised over a day) for day i, hour h, clustered in s.

A clustering-classification approach with different subsets of days is considered to infer the usage patterns not accounting for the weather dependence component. In a first step, the clustering technique is used to detect the patterns from a subset of the daily load curves when low, or even null, weather dependence is expected (during March, April, May, September, October, and November). Then, in a second step, a classification of the rest of the daily load curves is made using the clustering model obtained in the first stage.

The results of the clustering technique over the residential customers with tariff 2.0A in the case study area are depicted in **Figure 9**:

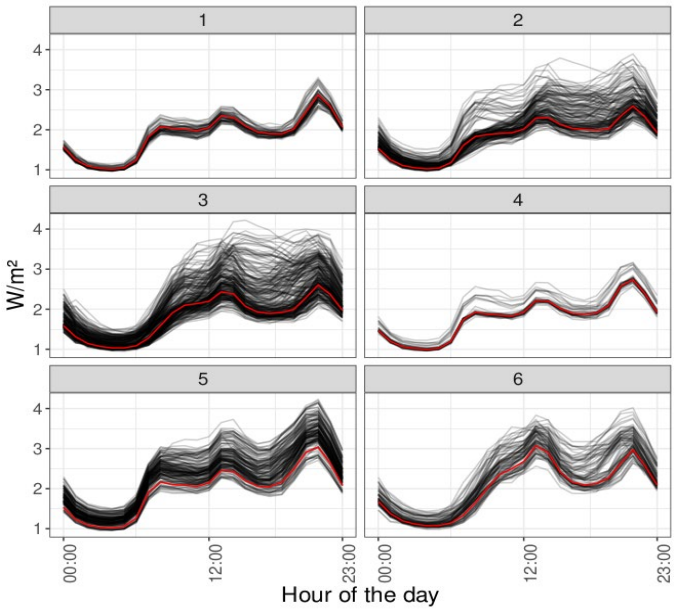
Figure 9. Clustering of the daily load curves, only using days that are presumably not affected by weather conditions. These six profiles represent the usage patterns of the case study



Source: CIMNE - BEE Group own elaboration

The red curves are the usage patterns, and the black curves are the actual daily load curves used during the training phase of the clustering procedure. Subsequently, using the same representation, the results of the classification stage are depicted (see **Figure 10**), where all the daily load curves, including winter and summer periods, are considered. As seen, the weather conditions' influence tends to increase energy consumption in certain usage patterns. However, in all cases, they maintain the relative shape in terms of hourly relative consumption concerning the overall daily consumption.

Figure 10. Classification of the complete series using the representative usage patterns detected with the clustering technique



Source: CIMNE - BEE Group own elaboration

2.3.4 Data-driven model

The technique used to characterise the electricity consumption consists of a penalised multiple linear regression model. The terms of this model are explained more in detail in the following subsections. However, in essence, the consumption is decomposed into multiple parts: the usage patterns estimated with the previous clustering-classification technique; the calendar features, which allow modelling the hourly and weekly baseload patterns; and the weather features, which enable to estimate the increase in consumption when severe weather conditions occur.

The formula considered for the penalised regression model is:

$$Q_t^e = (B_t \times s_t) + (H_t \times dh_t) + (C_t \times dh_t) + \varepsilon_t$$

Where Q^e_t is the electricity consumption at instant *t;* B_t is the baseload terms interacting with the usage patterns (s_t), H_t and C_t are the weather dependence terms, during heating and cooling periods, respectively, and interacting with the hour of the day (dh_t). Lastly, $\boldsymbol{\varepsilon}_t$ is the error term of the model with $\boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{\sigma}^2)$.

2.3.4.1 Baseload terms

Normally, the baseload component is one of the most significant parts of electricity consumption. The formal definition of baseload consumption consists of the minimum level of demand on an electrical grid over a span of time. However, in the framework of this methodology, it is understood as hourly consumption with no weather dependence at all. Hence, the baseload component only depends on the representative usage pattern and the calendar variables of a certain day.

Given the regression model presented, differences in consumption along the week and the day are considered. For both of them, a Fourier series describing the weekly and daily cycle was used. This decomposition transformation reduces the dimension of the fitting problem in cases where input variables are periodic.

The baseload terms are described in detail below:

$$B_t = \omega_b + S_{N_d}(p_t^d) + S_{N_w}(p_t^w)$$

$$S_{N_d}(p_t^d) = \sum_{n=1}^{N_d} \omega_{b,d,n,cos} \cos(2\pi n p_t^d) + \omega_{b,d,n,sin} \sin(2\pi n p_t^d) \qquad p_t^d = \frac{dh_t}{24}$$

$$S_{N_w}(p_t^w) = \sum_{n=1}^{N_w} \omega_{b,w,n,cos} \cos(2\pi n p_t^w) + \omega_{b,w,n,sin} \sin(2\pi n p_t^w) \qquad p_t^w = \frac{wh_t}{168}$$

Where ω_b is the linear intercept ; $S_{Nd}(p^d_t)$ and $S_{Nw}(p^w_t)$ are the Fourier series of the daily and weekly cycles, where $\omega_{b,w,n,cos}$, $\omega_{b,w,n,sin}$, $\omega_{b,w,n,cos}$ and $\omega_{b,w,n,sin}$ are the coefficients estimated within the regression model, N_d and N_w are the number of harmonics of both series, and finally, p^d_t and p^w_t are the relative part the day or the week at instant *t*. The dh_t and wh_t variables regard the hour of the day (dh_t), or the week (wh_t), at instant *t*.

The advantage of using the Fourier series is that it avoids the usage of an excessive number of dummy variables which would require the fit of all-possible combinations (24 + 168 dummy variables, in the case of fitting the regression model using an hourly-frequency dataset, multiplied by the number of usage patterns detected in previous steps). This transformation reduces the fitting problem to the number of harmonics considered (normally, between 3 and 5 harmonics per cycle), which are enough to infer the underlying correlation between the electricity consumption and the seasonal cycle in study without a considerable loss of information. Additionally, an interesting feature of the Fourier series transformations is that, in some sense, it coerces the regression to maintain a relationship between closer parts of the cycle and between the beginning and the end of the cycle itself.

2.3.4.2 Weather dependence components

Besides the baseload terms, heating and cooling dependent components account for the consumption made due to the weather conditions of the building location, the energy performance and characteristics of the buildings, and the operating behaviour of the HVAC systems.

This component estimates the increase in consumption due to weather severity. It is a key factor in understanding electricity consumption and infer characteristics of how the reference building/dwelling in a certain zone is composed and operated.

Ideally, one of the most interesting building characteristics that could be inferred using this type of modelling is the building envelope's Heat Transfer Coefficient (HTC). This coefficient highly depends on the considerations made during its estimation. For instance, depending on the consideration or not of certain components, such as ventilation or air leakage factors, the HTC should be differently understood. If ventilation and air infiltration features, the envelope represents all the surrounding surfaces of the building in contact with the outdoors, ground or other buildings. If it is not included in the model, the envelope represents, additionally to the surrounding surfaces, the energy transfer due to ventilation and infiltrations.

Furthermore, to estimate HTC, some additional variables are needed, such as indoor temperatures or performance characteristics regarding the HVAC systems installed in the buildings. Without this additional information, it becomes nearly impossible to estimate the HTC. Thus, in the framework of this methodology, instead of characterising the HTC as a heat energy quantification, an estimation of electricity consumed over the baseload due to an increase in indoor-outdoor temperature differences is performed.

To do so, and considering that only the wind speed and the outdoor temperature are available, multiple-input transformations over these features are considered to account for the different interactions between the electricity consumption and the weather conditions.

The first transformation considers the temperature differences between a theoretical balance temperature and the actual outdoor temperature. The main reason is to overcome the non-linearities between the outdoor temperature and consumption. Furthermore, different balance temperatures are considered during the heating and cooling season and during multiple parts of the day. This feature helps the model to characterise better certain situations. For instance, regions that require heating and cooling needs at the same time, or significant differences of weather dependence along the day. The increase in consumption due to an increase of this feature tends to be more related to ventilation systems without heat recovery units or window operations. Physically, it could be understood as colder or hotter outdoor air, compared to indoor air, enters the building, and the HVAC systems need to react to this fact (increase the consumption).

The second transformation uses the wind speed product and the theoretical temperature difference obtained by the first transformation to correlate consumption and the air infiltrations of the building. Those are the accidental introduction of outside air into a building, typically through cracks in the building envelope, doors, windows, and chimneys. It is caused by wind, negative pressurisation of the building, and air buoyancy forces known commonly as the stack effect. Normally, the higher is the wind speed and the temperature difference, the more energy consumption is made due to air infiltrations. Making a similar interpretation to the first transformation feature, HVAC systems need to increase the consumption to maintain the normal indoor thermal conditions.

Finally, another useful transformation is the consideration of low pass filters in the inputs of the model. Due to building inertia and heat transfer through the envelope, the indoor temperature of buildings does not react instantly to changes in the outdoor temperature. Then, to linearise the correlation between energy losses and energy consumption, a first-order low pass filter of the outdoor temperature T^o with a certain α parameter is considered. This tuned temperature is called $T^{o,lp}$, and, afterwards, it is transformed using the same differential process used in the first transformation. The low pass filter allows retaining the slow undisturbed variations (signals with a low frequency), while the fast variations are damped (filtered). It allows transforming the temperature used as input in the models into a variable that better represents the system's dynamics to enhance the fit of the model. It assumes that the dynamics of the buildings can be described by lumped parameter RC models. This assumption means that the response in consumption due to envelope energy transfers can be modelled as a first-order low pass filter.

To summarise the following terms are considered:

$$H_t = \omega^+_{h,lp} T^{h,lp}_t + \omega^+_h T^h_t + \omega^+_{ah} A^h_t$$

$$C_t = \omega^+_{c,lp} T^{c,lp}_t + \omega^+_c T^c_t + \omega^+_{ac} A^c_t$$

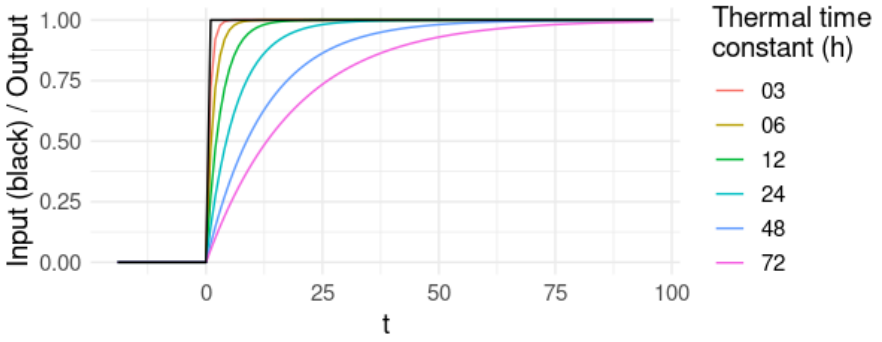
$$T^{h,lp}_t = (T^{bal,c}_{dh_t} - T^{o,lp}_t) d_{s_t}$$
$$T^h_t = (T^{bal,h}_{dh_t} - T^o_t) d_{s_t}$$
$$A^h_t = W^s_t T^h_t d_{s_t}$$
$$T^{o,lp}_t = \begin{cases} \alpha T^o_t & \text{if } t = 0, \\ \alpha T^o_t + (1-\alpha) T^{o,lp}_{t-1} & \text{if } t > 0. \end{cases}$$
$$d_{s_t} = \begin{cases} 1 & \text{if weather dependence in } s_t, \\ 0 & \text{if no weather dependence in } s_t. \end{cases}$$

$$T^{c,lp}_t = (T^{o,lp}_t - T^{bal,c}_{dh_t}) d_{s_t}$$
$$T^c_t = (T^o_t - T^{bal,c}_{dh_t}) d_{s_t}$$
$$A^c_t = W^s_t T^c_t d_{s_t}$$
$$\alpha = 1 - e^{-t_{sampling}/(2\pi\tau/24)}$$

Where: $\omega^+_{h,lp}$ is the always-positive linear coefficient for the heating dependent term that considers the thermal inertia of the reference building ($T^{h,lp}_t$), which is calculated as the difference between balance heating temperature ($T^{bal,h}_{dht}$) at the part of the day (dh_t) and the low-pass filtered outdoor temperature ($T^{o,lp}_t$) at instant t; ω^+_h is the always-positive linear coefficient for the raw heating dependent term(T^h_t), which is usually related to ventilation heat losses, and it is calculated as the difference between balance heating temperature ($T^{bal,h}_{dht}$) at the part of the day (dh_t) and the raw outdoor temperature (T^o_t) at instant t; ω^+_{ah} is the always-positive linear coefficient for the heat losses due to air infiltrations (A^h_t), which is the wind speed (W^s_t) multiplied by the raw heating dependent term(T^h_t) at instant t; $\omega^+_{c,lp}$ is the always-positive linear coefficient for the cooling dependent term that considers the thermal inertia of the reference building ($T^{c,lp}_t$), which is calculated as the absolute difference between balance cooling temperature ($T^{bal,c}_{dht}$) at the part of the day (dh_t) and the low-pass filtered outdoor temperature ($T^{o,lp}_t$) at instant t; ω^+_c is the always-positive linear coefficient for the raw cooling dependent term(T^c_t), which is usually related to ventilation heat gains, and it is calculated as the difference between balance cooling temperature ($T^{bal,c}_{dht}$) at the part of the day (dh_t) and the raw outdoor temperature (T^o_t) at instant t; ω^+_{ac} is the always-positive linear coefficient for the heat losses due to air infiltrations (A^c_t), which is the wind speed (W^s_t) multiplied by the raw cooling dependent term(T^c_t) at instant t.

In addition, the α value of the outdoor temperature low-pass filter depends on the $t_{sampling}$, which is the number of measures per hour, and the τ thermal time constant, which is the number of hours needed by the reference building to react over a certain change in outdoor temperature. **Figure 11** provides an example of different τ (e.g. thermal time constant) values and the step function between input (e.g. raw temperature) and output (e.g. low-pass filtered outdoor temperature).

Figure 11. Example of a first-order low pass filter depending on a set of different time constants



Source: CIMNE - BEE Group own elaboration

Last but not least, all the temperature differentials and air leakage terms are multiplied by a dummy variable which coerces weather dependence terms to 0 if a certain usage pattern has no weather dependence (d_{st}).

2.3.4.3 Impact of holiday seasonality's

After the first tests of the implementation, it was noticeable that the influence of holidays tends to generate significant changepoints in electricity consumption for certain regions, sectors and periods of the year. In most cases, the holidays periods were national holidays, Fridays or Mondays between national holidays and weekends, winter and summer weekends, and summer vacations. Nonetheless, it was difficult to find a feature that correlates linearly to the holidays component of the electricity consumption due to the different local casuistic of every region along the year. For instance, some of the features that could be used are number of tourists, second homes occupancy or number of hotel bookings at the postal code level and daily frequency. After successive attempts, this information was impossible to find at the desired aggregation levels.

Therefore, another strategy is considered in the final implementation. The data-driven characterisation model is fitted using only those days that are not suitable to be holidays. Then, the whole period is predicted using the trained model. The residuals between the actual and predicted data during the holidays period are considered the holiday's component. In addition, this holidays dependence component is estimated only when a difference of at least 20% was detected between the RMSE of the holidays / non-holidays period.

2.3.4.4 Impact of Covid-19 lockdown

The Covid-19 Spanish lockdown, from 15 March to 21 June 2020, affected the energy consumption drastically either in residential, industrial or public sectors. Changes in business activities, user behaviour and building occupancy caused this situation. For the case study presented, the period analysed depends on the availability of electricity consumption data for each postcode. In general, the evaluated period comprised from mid-2018 to mid-2020. Thus, the data used to validate the characterisation methodology was fully affected by this lockdown period.

To quantify the decrease or increase in consumption due to the lockdown, a set of terms are introduced into the regression model, which adds an interaction of the lockdown period to the baseload terms and a set of re-adjusted weather dependence coefficients during the period.

Another consideration made during the period was that the holidays' affectation must be fixed to zero, as people should stay at home during those periods, except for specific purposes. Hence, the affectation of holidays, seasons or weekends were residuals compared to the normal behaviour.

2.3.4.5 Training of the model

The electricity time series considered during the training phase changes slightly depending on the economic sector considered. The reason behind this discretisation is to consider the most representative area factor for each economic sector, as the outcomes of the characterisation are afterwards compared among different regions, and the built area normalisation becomes a key factor in the assessment of the energy performance of buildings. The ratios considered for each location and existing tariffs are the following:

— Residential sector:

$$Q^e = \frac{\text{Total consumption}^{residential}}{\text{number of contracts}^{residential} \times \text{average dwelling area}} [W/m^2]$$

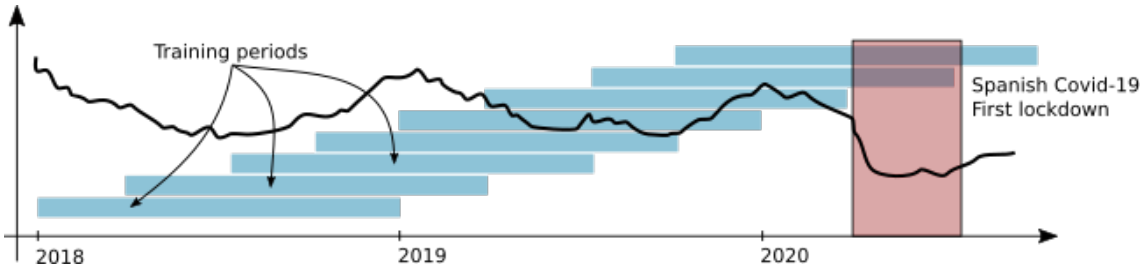
— Industrial / Agriculture / Offices / Retail sector:

$$Q^e = \frac{\text{Total consumption}^{sector}}{\text{number of contracts}^{sector} \times \text{average building area}^{sector}} [W/m^2]$$

The model's training is recursively performed every three months over a one-year window (see **Figure 12**). This fact provides information on how the reference building is evolving in time. So, the characterisation coefficients

become, in some sense, time-variant. The original hourly frequency of the input time series is resampled to 4 hours to decrease the computational time.

Figure 12. Model training periods to characterise the evolution in time of the dependencies



Source: CIMNE - BEE Group own elaboration

Regarding the estimation of the unknown terms, most of them are inferred through the maximum likelihood technique implemented in the *penalised* function of the R package *Penalised* [12], where the whole regression formula is estimated. However, several coefficients cannot be solved using this methodology, as they are variables that transform the model inputs themselves. Examples are the thermal time constant of the reference building, the number of harmonics of the Fourier series, or the balance temperatures, among others.

Known terms and time series: Q^e, s, p^d, p^w, dh, wh, T^o, W^s and $t_{sampling}$.

Unknown fixed terms: $\tau(*)$, N_d and N_w.

Unknown terms for each usage pattern: ω_b, $d_s(*)$, $\omega_{b,d,n,sin}$, $\omega_{b,d,n,cos}$, $\omega_{b,w,n,sin}$ and $\omega_{b,w,n,cos}$.

Unknown terms for each day part: $\omega^+_{h,lp}$, ω^+_h, ω^+_{ah} ,$\omega^+_{c,lp}$, ω^+_c, ω^+_{ac}, $T^{bal,h}_{dh}(*)$ and $T^{bal,c}_{dh}(*)$.

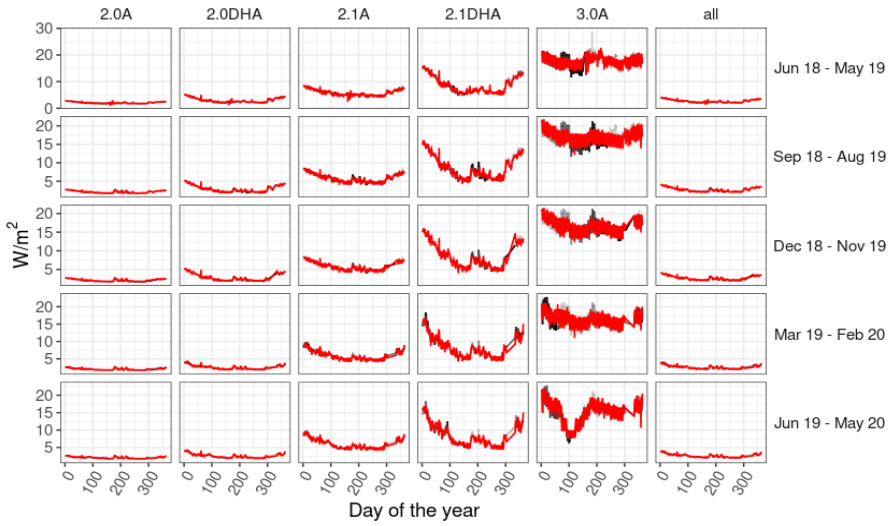
(*) Estimated using a genetic algorithm optimizer

The optimisation of these coefficients is made using a Genetic Algorithm that iterates and evolves chromosomes (in this case, the coefficients' values to optimise), minimising a cost function. This function is the RMSE of the predicted consumption versus the real consumption data.

2.3.5 Case study results

In this section, the characterisation results of the case study of the 25006 postal code are explained in detail. These results exemplify some of the insights that can be inferred using this methodology. Firstly, **Figure 13** shows how prediction (red) fits the real data (black).

Figure 13. Accuracy of the characterisation model over distinct periods and tariffs



Source: CIMNE - BEE Group own elaboration

The accuracy of the models for each of the tariffs and evaluation periods are detailed in the next table:

$$MAPE = \sum_{t=1}^{N} \left| \frac{observed_t - predicted_t}{observed_t} \right| \times \frac{100}{N}$$

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n} (X_{obs,i} - X_{pred,i})^2}}{n}$$

$$CV(RMSE) = \frac{RMSE}{\overline{X_{obs}}}$$

Table 2. Mean Average Percentage Error (MAPE) over distinct periods and tariffs

MAPE (%)	2.0A	2.0DHA	2.1A	2.1DHA	3.0A	all
June 2018 - May 2019	4,52	7,05	5,78	8,28	7,03	5,33
Sept. 2018 - Aug. 2019	4,31	7,65	5,90	9,30	6,89	5,02
Dec. 2018 - Nov. 2019	4,18	6,25	5,56	8,36	5,92	4,73
Mar. 2019 - Feb. 2020	4,37	5,95	6,35	9,79	6,52	5,34
June 2019 - May 2020	4,15	5,32	5,57	8,69	7,35	4,77

Source: CIMNE - BEE Group own elaboration

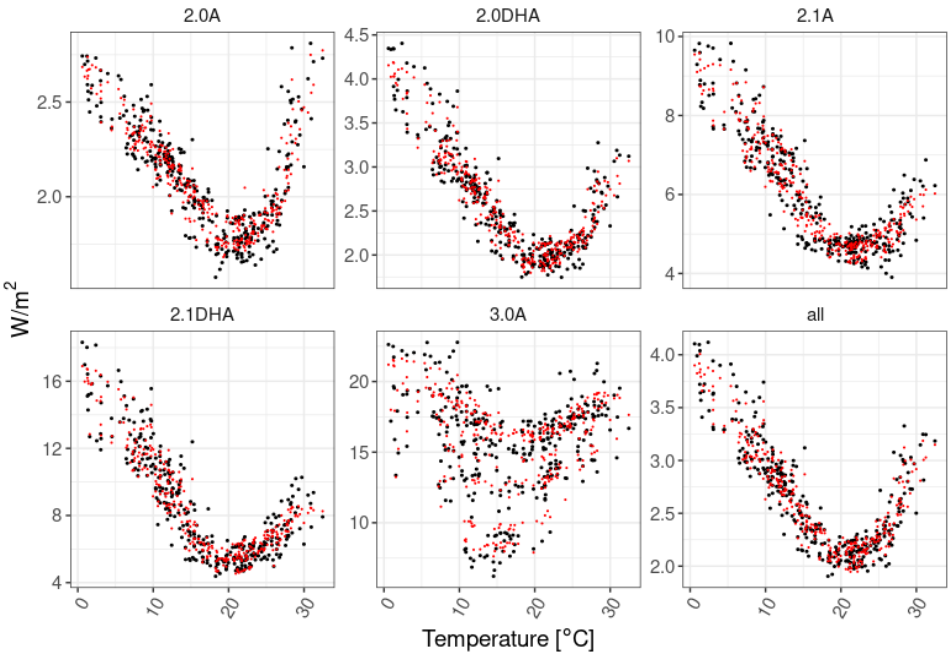
Table 3. Coefficient of Variation (CV) of the RMSE over distinct periods and tariffs

CVRMSE [%]	2.0A	2.0DHA	2.1A	2.1DHA	3.0A	all
June 2018 - May 2019	5,75	8,53	7,34	9,99	8,94	6,45
Sept. 2018 - Aug. 2019	5,68	9,08	7,56	10,68	8,55	6,55
Dec. 2018 - Nov. 2019	5,65	8,06	7,40	10,27	7,84	6,27
Mar. 2019 - Feb. 2020	5,95	7,73	8,25	12,40	8,61	7,03
June 2019 - May 2020	5,56	7,06	7,06	11,03	8,58	6,27

Source: CIMNE - BEE Group own elaboration

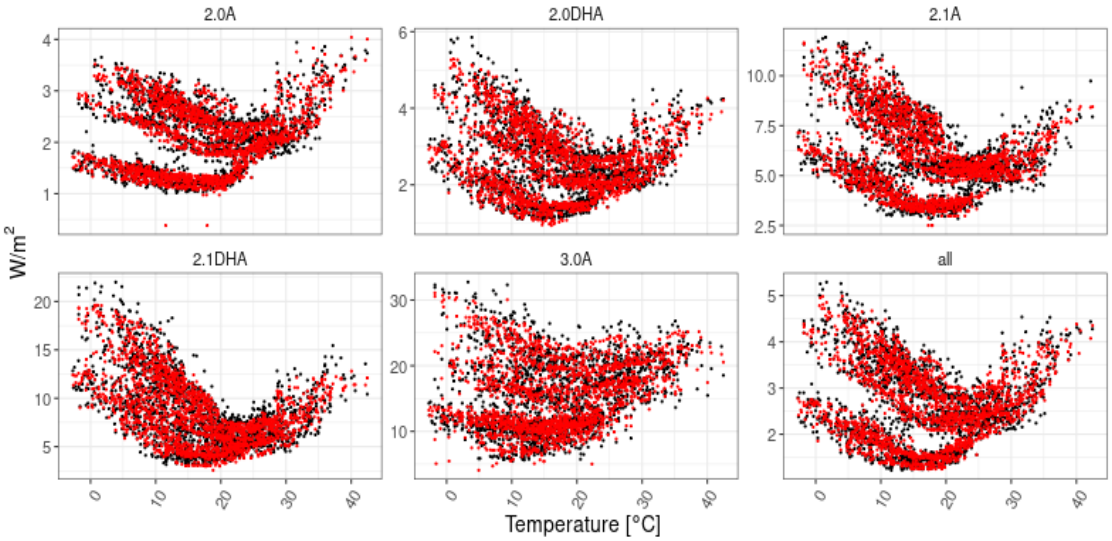
For both of the selected metrics, the accuracy levels reach a good level for characterisation purposes. Another interesting correlation between real observations and predictions from the data-driven models is to compare the energy signatures, either using a daily (see **Figure 14**) or a 4-hours (see **Figure 15**) resampling. In both cases, it has been proved that the predictions follow the main tendency and the variance of the consumption - temperature correlation.

Figure 14. Predicted daily energy signature versus actual data



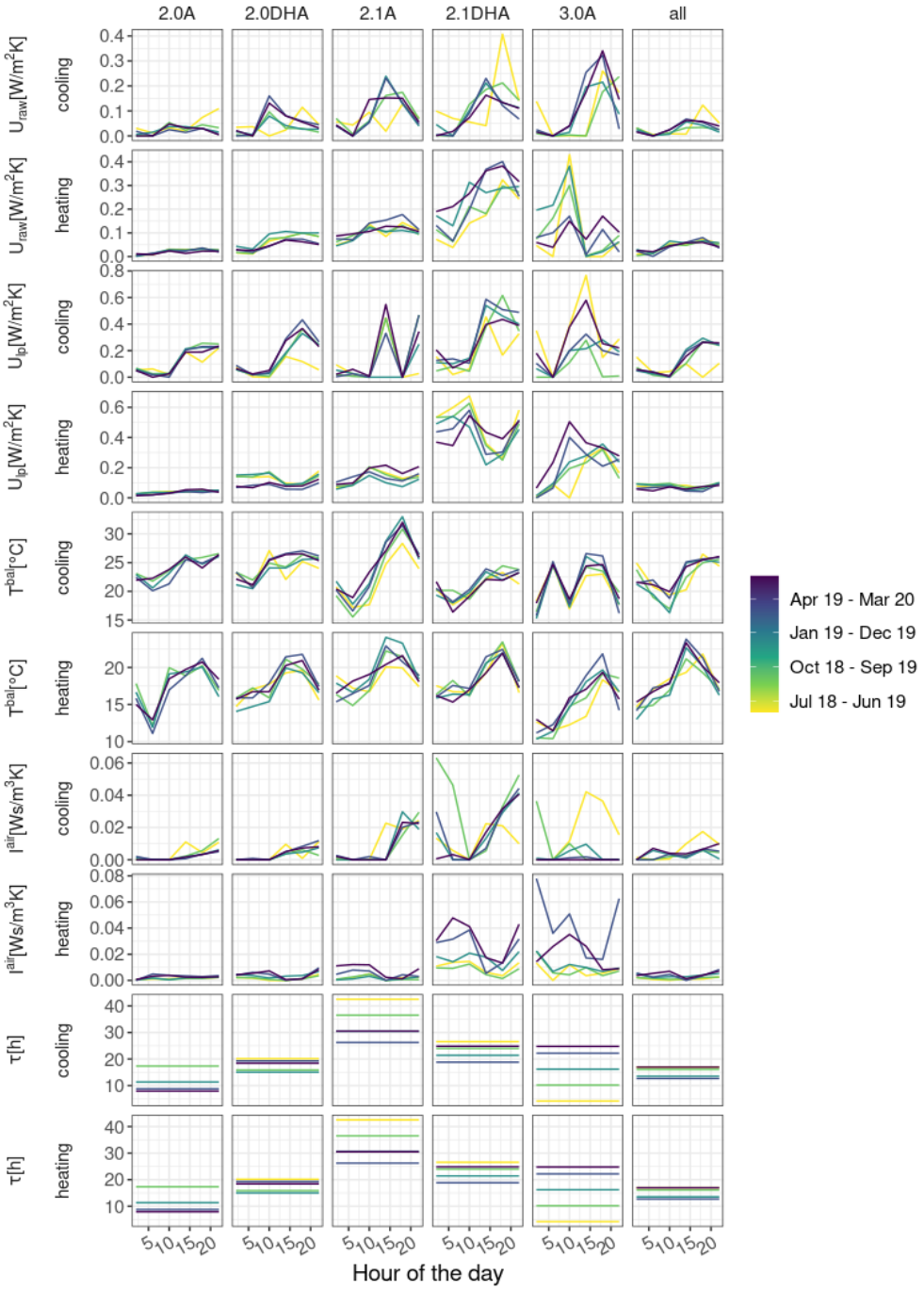
Source: CIMNE - BEE Group own elaboration

Figure 15. Predicted 4-hourly aggregated energy signature versus actual data



Source: CIMNE - BEE Group own elaboration

Figure 16. Weather-dependent characterisation parameters of the model

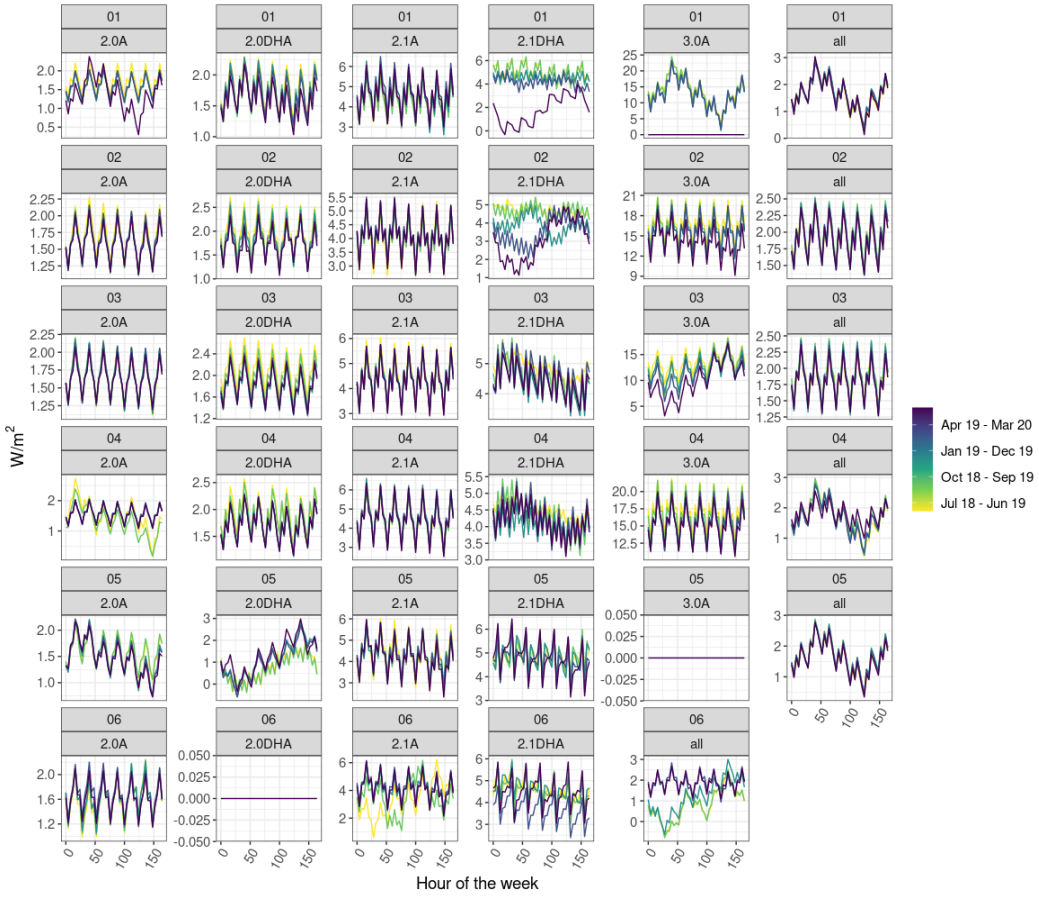


Source: CIMNE - BEE Group own elaboration

The weather-related coefficients are depicted in **Figure 16**. Dark blue lines correspond to the characterisation during June 2019 - May 2020 and the yellow ones to July 2018 - June 2019. In the Y-axis, the different weather dependence coefficients in heating and cooling modes are depicted. U_{raw} heating values are the ω^{+}_{h} model coefficients depending dh_t (hour of the day), U_{lp} heating values are the $\omega^{+}_{h,lp}$ model coefficients depending the dh_t , I^{air} heating values are the ω^{+}_{ah} model coefficients depending the dh_t , T^{bal} heating values are the heating balance temperature depending the dh_t , τ heating is the thermal time constant during the heating season, U_{raw} cooling values are the ω^{+}_{c} model coefficients depending dh_t , U_{lp} cooling values are the $\omega^{+}_{c,lp}$ model coefficients depending the dh_t , I^{air} cooling values are the ω^{+}_{ac} model coefficients depending the dh_t , T^{bal} cooling values are the cooling balance temperature depending the dh_t , and τ cooling is the thermal time constant during the cooling season.

It can be seen that the coefficients across different tariffs vary largely, and tend to be higher, the more consumption is made by the tariff customers. This is a normal effect, as customers with 2.1 and 3.0 tariffs, tend to have more electricity or HVAC equipment on their dwellings or buildings. One of the most interesting outcomes is that this characterisation methodology accounts for the heating and cooling dependencies on different parts of the day. Hence, it could be inferred which part of the day the reference building or dwelling responds with more emphasis against the weather conditions. The balance temperature helps to understand the most common HVAC operation schedule during a typical day or how people or energy managers set the setpoints of their thermostats. Additionally, differences in the thermal time constant show differences in building's envelope characteristics among different tariffs. At a glance, it seems that 2.1 and 3.0 tariffs, in the residential economic sector are related to higher thermal inertia buildings or better insulation.

Figure 17. Baseload-dependent characterisation parameters of the model

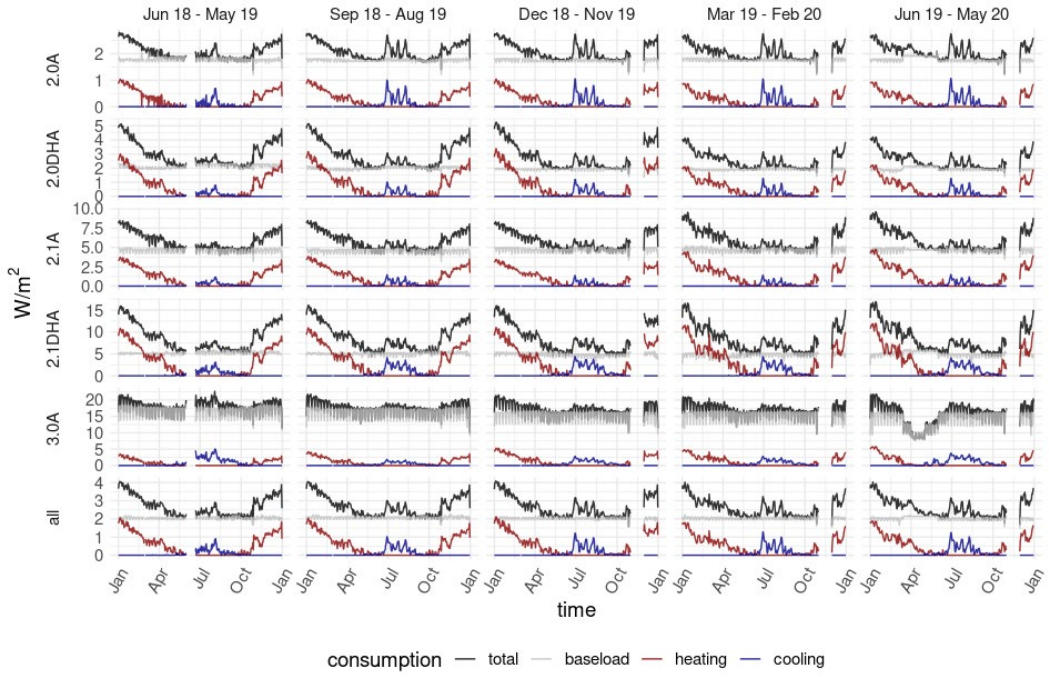


Source: CIMNE - BEE Group own elaboration

Regarding the baseload characterisation, the daily and weekly profiles for each usage pattern and tariffs are depicted in a similar way to the weather dependence coefficients. **Figure 17** shows the evolution of the baseload along the period June 2018 - May 2019 to June 2019 - May 2020. The dark blue lines correspond to the most updated coefficients, and their remarkable differences are produced, in some cases, along the day of the week; even the daily profile is quite similar in terms of relative consumption between different parts of the day.

In summary, using the regression model obtained, the decomposition of the three components (baseload, heating, cooling) is made for the whole period of data affected by each of the evaluation periods (June 2018 - May 2019 to June 2019 - May 2020). In the web application, the results of this disaggregation are much better represented using interactive plots. However, to show the results in a paper format, **Figure 18** represents the daily disaggregation and the total consumption. For the sake of comparison between different periods, all the X-axis represent the months from January to December.

Figure 18. Daily electricity disaggregation results over distinct periods and tariffs



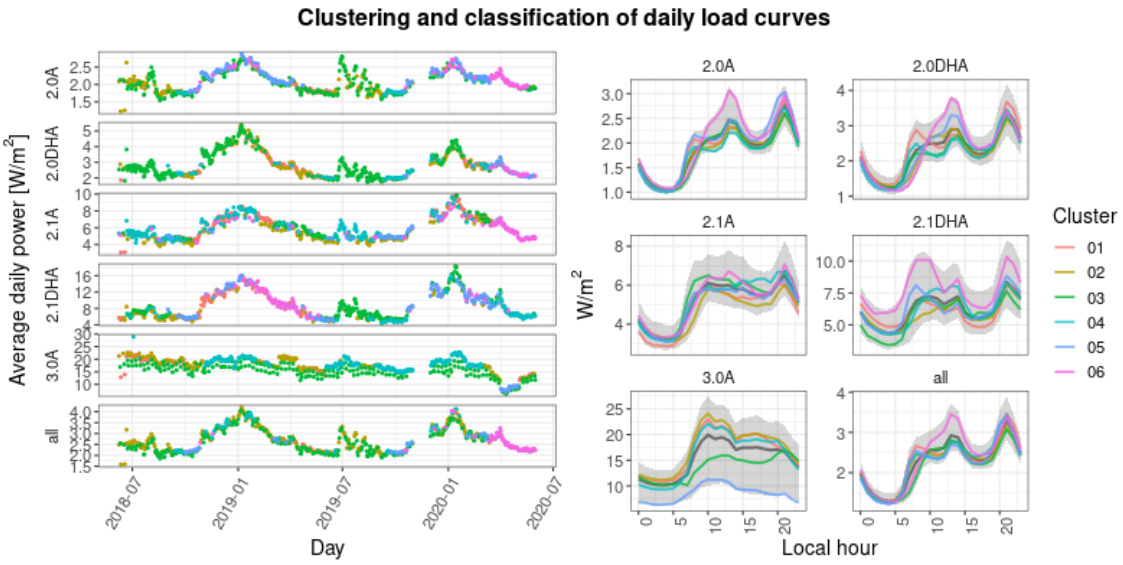
Source: CIMNE - BEE Group own elaboration

It is noted that, in all the cases, the significance of the baseload consumption is much higher than the weather dependence components. Also, stands out the high variance in the baseload component in tariff 3.0A, which corresponds to the weekdays-weekends variation. Another detail that can be seen in this plot is the impact of the Covid-19 lockdown in Spain during the months from March to May of the last evaluation period, especially in the case of tariff 3.0A, where allegedly some business buildings/dwellings are integrated in the residential sector subset of the Datadis database. The evolution of the heating and cooling components through the year seems to fulfil the normal behaviour during a natural year, considering the total consumption series and the climate data of the case study area. However, it is noted that the reference building of tariff 2.1DHA has the major impact in terms of heating dependency. So, it can be interpreted that customers with this tariff have more electricity resourced heating systems compared to the customers with other tariffs.

2.4 KPIs of the characterisation

Once the characterisation model is technically fitted, a set of KPIs is defined to make comparisons along with different areas, even when certain conditions of those areas differ widely from the point of view of the type of users, weather conditions, or building characteristics. To do so, simple units and plots were chosen to represent the model results.

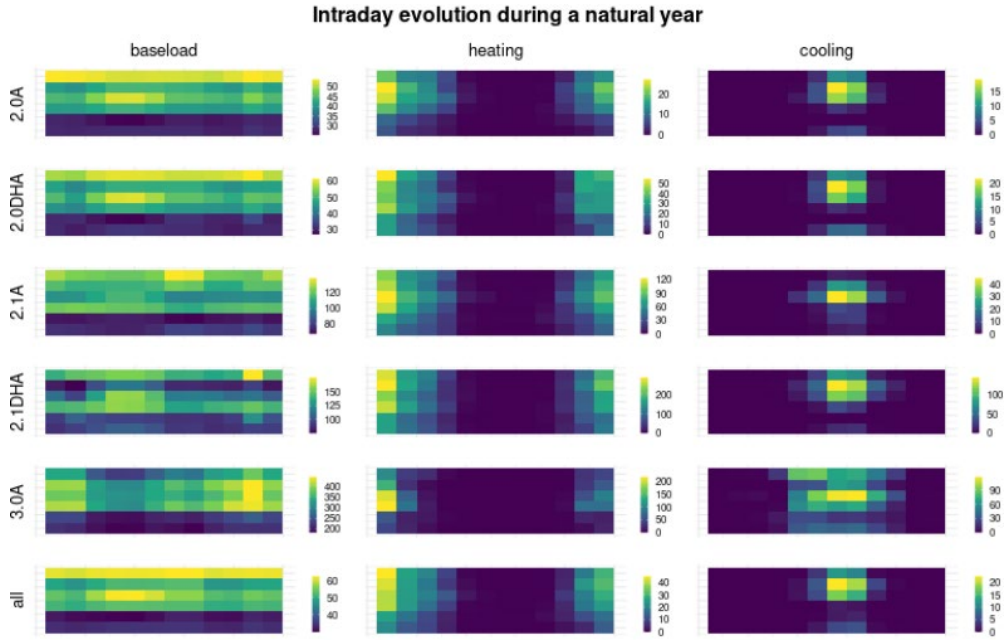
Figure 19. Usage patterns detected over distinct tariffs



Source: CIMNE - BEE Group own elaboration

The results of the clustering and classification of the usage patterns are depicted using **Figure 19**. In the right pane, the different usage patterns in multiple colours are displayed. In grey, the interval of daily load curves at confidence 95% is shown. In the left pane, the daily classification is represented, and it can be observed that some patterns have continuity in time. Hence, they tend to evolve over time, depending on certain conditions that interact with energy consumption. These conditions are related to the weather, part of the year, holiday seasons and other unknown variables.

Figure 20. Intraday summarised electricity disaggregation results over a natural year and distinct tariffs



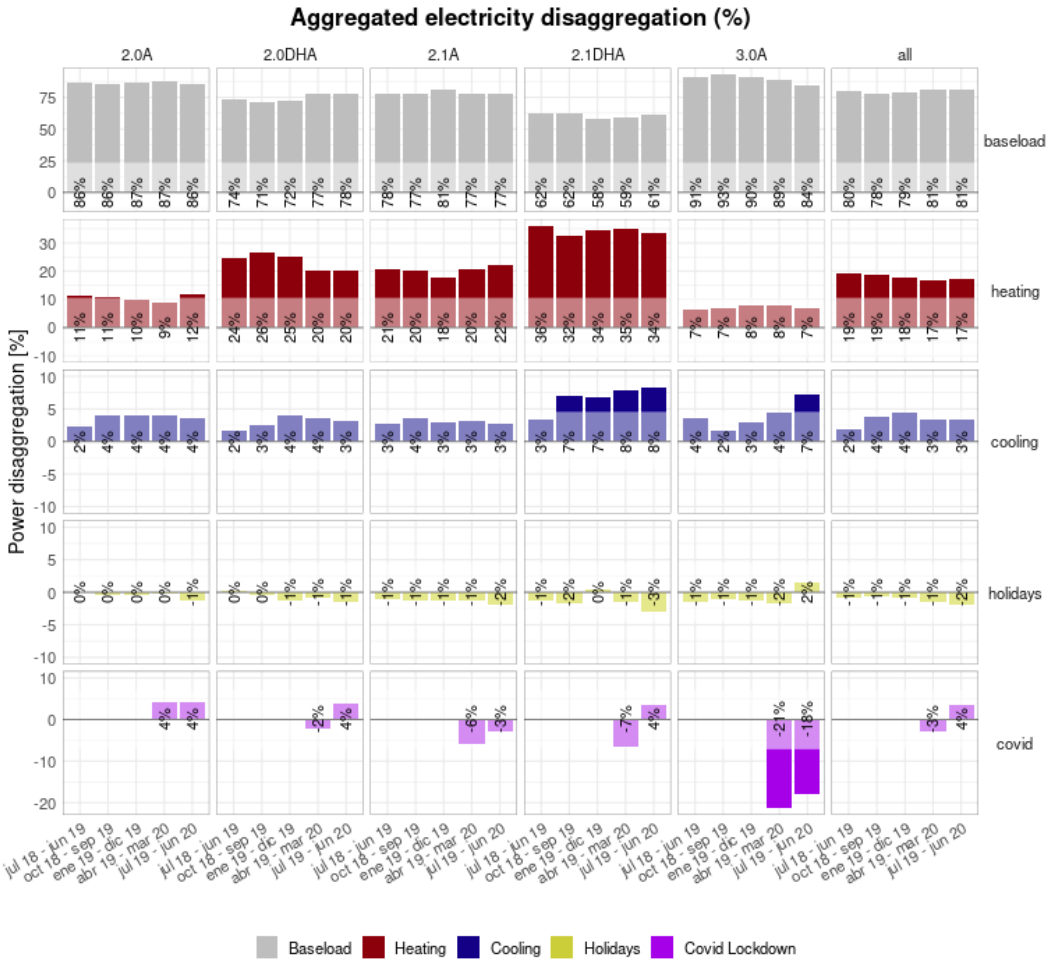
Source: CIMNE - BEE Group own elaboration

The heat map above uses the most updated characterisation model (Trained with data from July 2019 to June 2020) to show the average kWh/year contribution of each electricity component by tariff through a natural year (X-axis, each step is one month) and the different parts of the day (Y-axis, each step are four hours).

It can be seen that in the case of baseload, it seems that during the Covid-19, it has been incremented by about 20% during the step 12 to 16h. This can be related to more people in their homes cooking during lunchtime. On the contrary, 3.0A customers decrease their consumption drastically during those months. The heating and cooling components can be observed in the different intraday dependencies along different tariffs and months of the year (see **Figure 20**). Maybe, again, the 3.0A customers clearly behave significantly different in terms of cooling dependency compared to customers with other tariffs.

Additionally, to increase the understandability of the distribution between components and their evolution in time, **Figure 21** represents the relative disaggregation on a natural year basis between the baseload, heating and cooling components, and the affectation of holidays and Covid-19 lockdown on the total consumption.

Figure 21. Yearly-aggregated relative subcomponents of the electricity consumption over distinct periods and tariffs



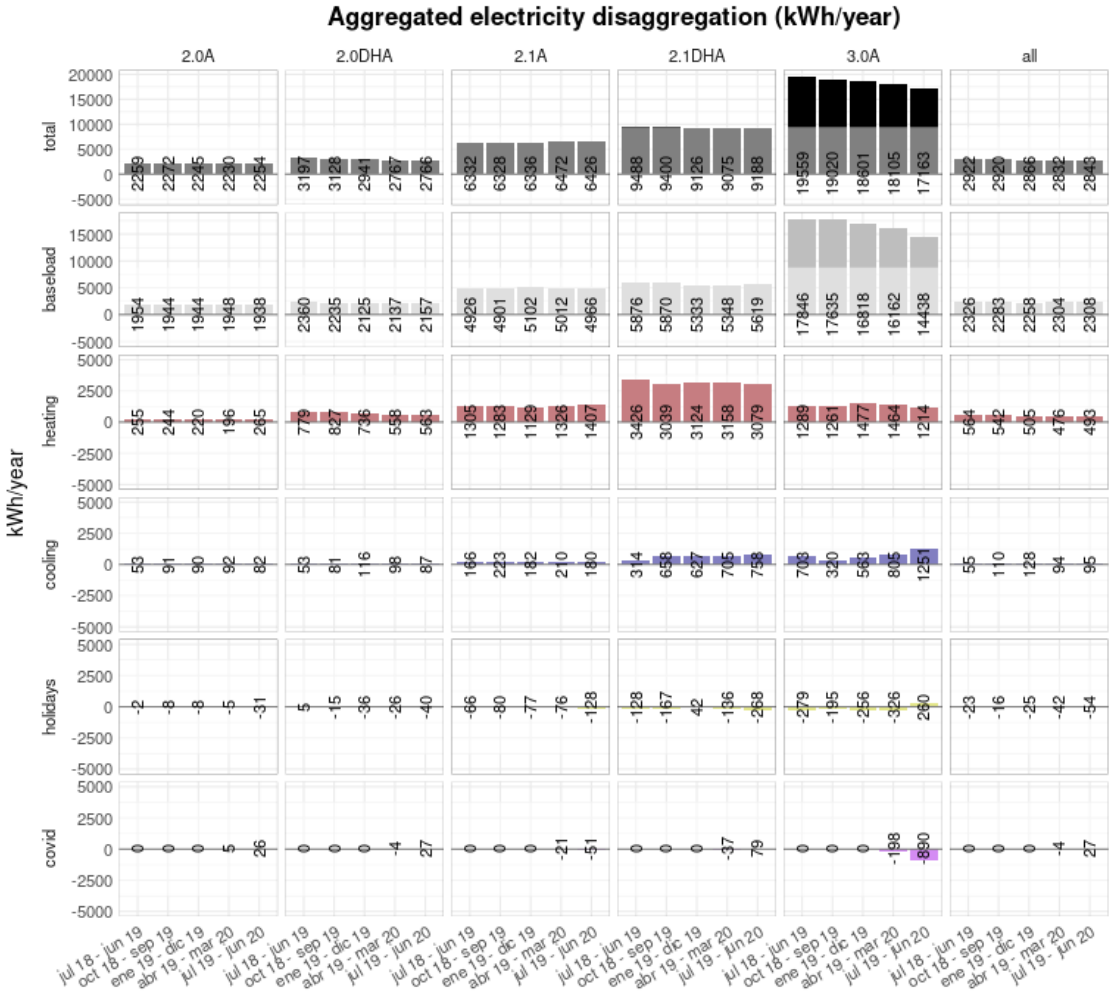
Source: CIMNE - BEE Group own elaboration

For instance, regarding tariff 2.0A and the first period July 2018 to Jun 2019: the baseload component represents approximately 86% of the total annual consumption, the heating component the 11%, the cooling component represents 2%, and the holidays periods do not contribute at all. In the Covid-19 lockdown percentage, the percentage means the relative amount of affectation during the lockdown period (15 Marc^{h} to 21 June 2020), not the total annual consumption.

Hence, in terms of conclusions taking a look at the evolution of subcomponents in time, as a general fact, the tendencies are very similar along time. However, significant differences can be detected between different

tariffs. This corresponds to the different users/building typologies that characterise each tariff, as explained in the electricity consumption data section.

Figure 22. Yearly-aggregated absolute subcomponents of the electricity consumption over distinct periods and tariffs



Source: CIMNE - BEE Group own elaboration

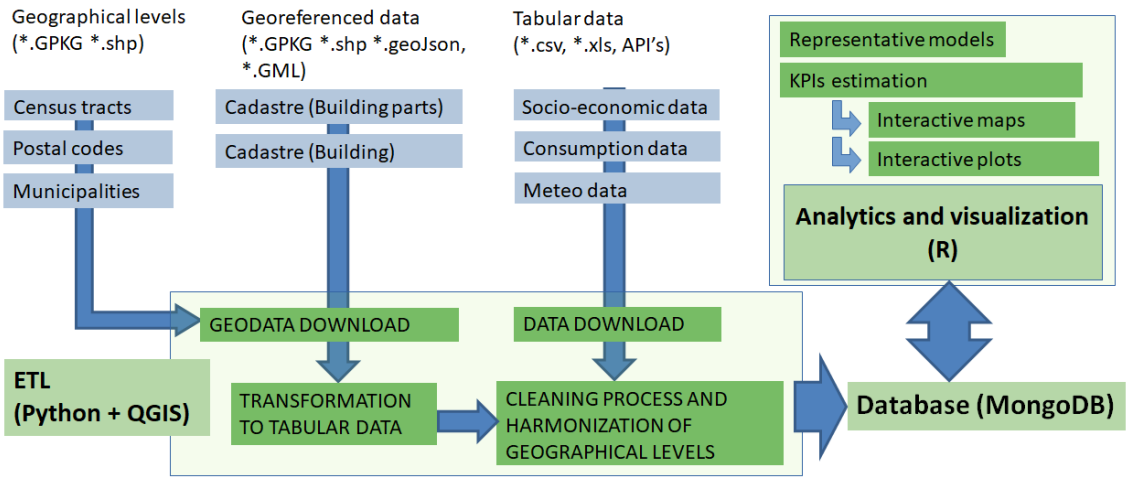
The absolute consumption contribution in kWh per natural year is illustrated in **Figure 22**. A decrease in total consumption in the 3.0A and 2.0DHA tariffs could be detected, especially the former, which is much more affected by the Covid-19 lockdown (approx. -20%). Then, in general, for the rest of the tariffs, the same amount of total consumption during the whole evaluation period is observed. And again, a similar conclusion to the ones detected using the relative disaggregation plot can be made.

2.5 Data integration

The implementation of this methodology consists of combining and analysing multiple layers of data, as shown in **Figure 23**. Considering that this information has heterogeneous characteristics, both in terms of frequency, geographical reference, and typology, one of the mandatory aspects of cross-analysis is the harmonisation of these layers. Specific aggregations and transformations are done for each input dataset. For instance, GML files of cadastre data are transformed to tabular data and aggregated to several geographical levels to be able to correlate cadastral information to socioeconomic conditions, electricity consumption and weather data.

Python 3.8 is used to extract, transform, and load data processes, using QGIS 3.10 as a backend to analyse geospatial data. Regarding the electricity characterisation model, it is implemented in R 4.1. All these scripts store the raw, intermediate and final results to a MongoDB 4 non-relational database.

Figure 23. General view of the data flow and the architecture of the software

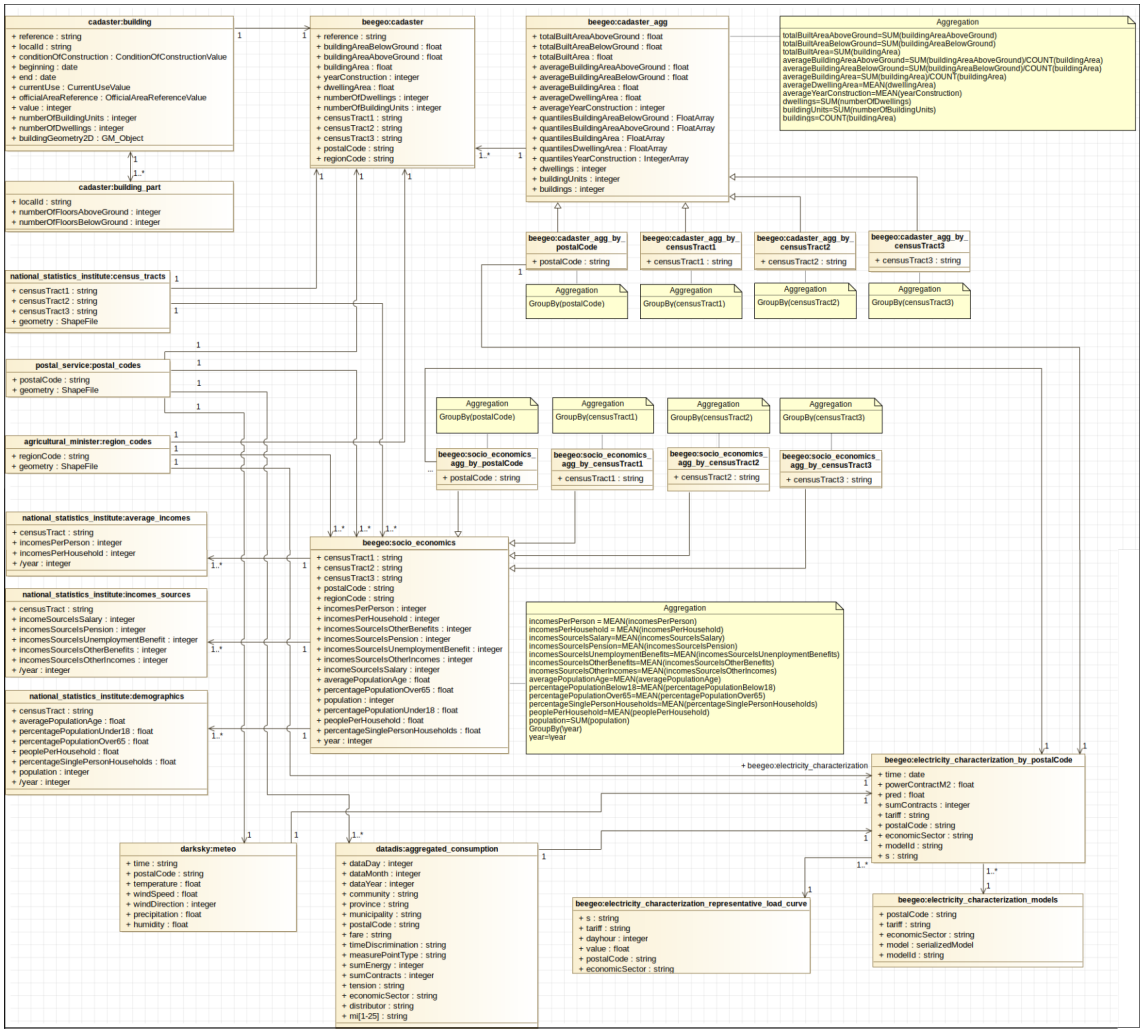


Source: CIMNE - BEE Group own elaboration

The relationships and transformations among the different databases are depicted in the UML model shown in **Figure 24**, where the classes are named using the following structure: "<name_of_the_provider>:<name_of_the_collection>". In the case of intermediate or final classes used by the data analytics backend or the frontend to visualise results, the provider's name is "beegeo". The calculations considered for the aggregations to higher geographical levels are explained in SQL format in yellow notes.

The implementation of this UML representation is made using a combination of open-source analytics and storage technologies that allow validating the methodology over the province of Lleida.

Figure 24. UML of the data model used in the Spanish implementation



Source: CIMNE - BEE Group own elaboration

3 Results of the validation

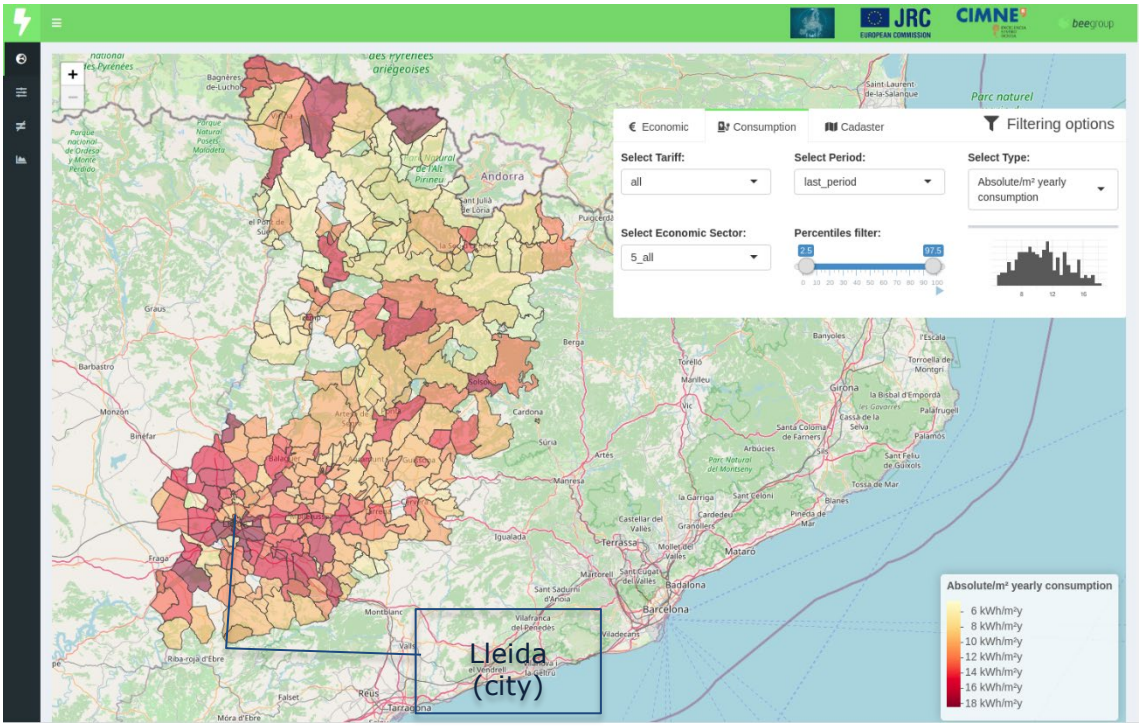
3.1 Data visualisation

The visualisation of the characterisation KPIs is made using an R Shiny frontend application, which has been developed on-purpose for this use case. In general, the data prompted into this web application is always read from the MongoDB database. However, some of the normalisation calculations are computed on-demand using the serialised characterisation models estimated in the analytics backend. The web application is mounted on Docker containers; hence, it should be horizontally scalable, which is an interesting feature for future deployment of the application, either for Spain or other EU countries. The time framework used in this validation concerns only data from the beginning of 2018 until June 2020, but the ETL processes are prepared to recursively obtain new data as soon as the inputs become available online.

To sum up, the web application is divided into four tabs: KPIs on a map, Characterisation, Benchmarking and KPIs correlation. In the following subsections, a detailed explanation of the objectives and the outcomes that can be reached in each tab is provided.

3.1.1 "KPIs on a map" tab

Figure 25. Web application - "KPIs on a map" tab



Source: CIMNE - BEE Group own elaboration

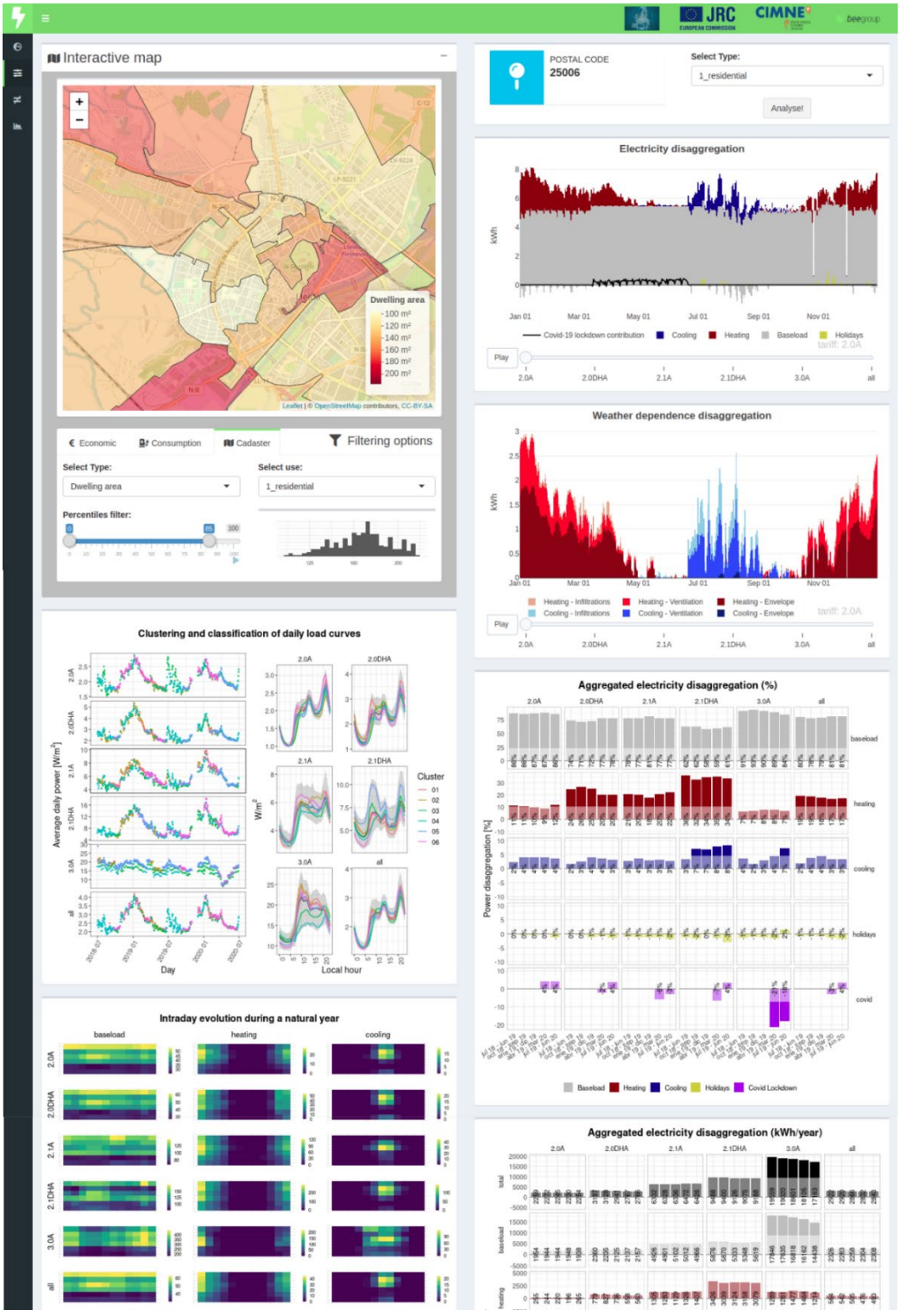
This is the home section of the web application. Its purpose is to give a clear and simple visualisation of all the estimated consumption KPIs, cadastre information and socioeconomic indicators. The visualisation could be filtered by tariffs, economic sectors, periods, percentiles ranges, and an interesting feature is a tiny histogram representing the distribution of values of the variable depicted on the map, especially important when outliers can generate useless colouring legends. The web application users can see at a glance several key factors for the characterisation of the whole region in analysis. For instance, in this case, the total annual consumption per dwelling area is depicted. Therefore, even if the KPI is normalised by area, there are large differences comparing the postal codes near Lleida (city) and the rest of the province. At 50 km around Lleida, the tendency is to expend much more electricity consumption compared to inner parts of the province. Nonetheless, there are several exceptions, especially in the Pyrenees.

To sum up, this section would be very useful for making a preliminary comparison along the region in analysis and taking the first conclusions depending on the desired application. For instance, a heat pumps dealer could

take a look at the heating and cooling components KPIs to understand where would be more suitable to make some advertising campaign or to prioritise the customer search areas of a salesperson during a certain period.

3.1.2 "Characterisation" tab

Figure 26. Web application – "Characterisation" tab (first part)



Source: CIMNE - BEE Group own elaboration

Figure 27. Web application – "Characterisation" tab (second part)

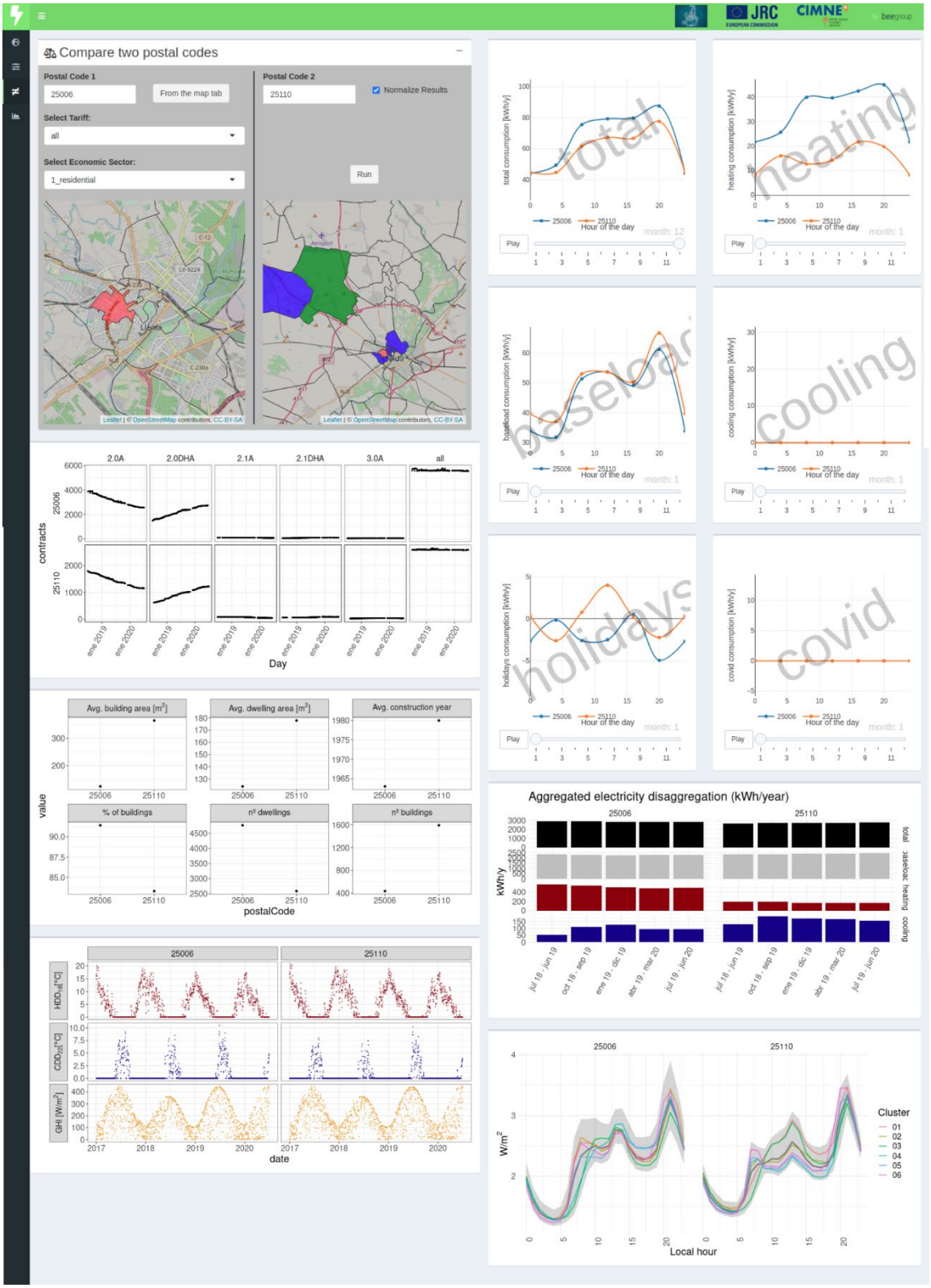
Source: CIMNE - BEE Group own elaboration

The characterisation section provides detailed information about electricity consumption usage on a single postal code. Several of the plots shown in this tab are interactive versions of the KPIs results explained in the section above, such as information about the model accuracy, the usage patterns detected and the disaggregation results in several time aggregation levels.

Using this tab, the potential web application user can go deeper into the detail of the most common electricity uses in a certain geographical area. For example, to understand the principal differences within tariffs over a certain economic sector and use this information to estimate the impact of a certain application, or to evaluate when some dependency occurs in time, or its evolution in time, which could indicate changes in the electric equipment, user behaviour or energy performance indicators of the building.

3.1.3 "Benchmarking" tab

Figure 28. Web application – "Benchmarking" tab



Source: CIMNE - BEE Group own elaboration

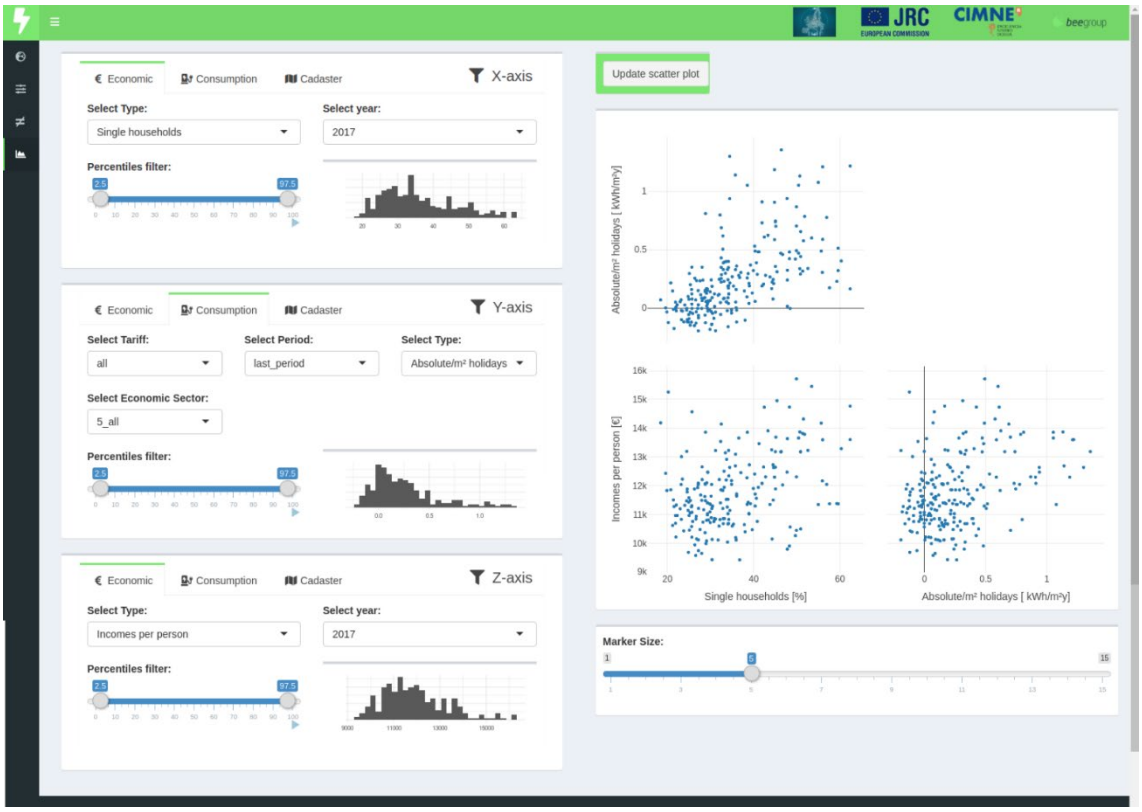
In this tab, the objective is to exploit the usage of the characterisation models to compare in detail two postal codes normalising the results by weather conditions and building/dwelling sizes. This normalisation procedure

means that the difference in electricity consumption should be caused by the energy performance of buildings or by the occupants HVAC operation schedule in the case of cooling and heating electricity consumption components. In parallel, intraday differences along a natural year between the baseload consumption, and the contribution of holidays periods and the Covid-19 lockdown period, are also depicted. In all these cases, the normalisation against the size of buildings/dwellings is also performed so that the benchmarking conclusions could be more related to differences in usage patterns or building occupancy.

To sum up, this section of the web application is a second derivative of the characterisation section. From a user point of view, this can help figure out the reason behind a significant change in electricity consumption. The normalising feature implemented in this section filters two of the most important factors that cause a change in electricity consumption: the size of buildings and the weather severity. Hence, more realistic comparisons could be made in terms of EPB and use trends between pairs of geographical areas in analysis.

3.1.4 "KPIs correlation" tab

Figure 29. Web application - "KPIs correlation" tab



Source: CIMNE - BEE Group own elaboration

All the KPIs represented in the map section can be cross-correlated in this tab of the web application. This application is quite helpful to understand tendencies and relations between the KPIs, providing a wider characterisation of the territory, to understand if the increase of a certain cadastre or socioeconomic indicator has a significant correlation to another energy consumption KPI. For instance, it could be inferred if there's a relation between holidays periods contribution to the energy consumption and average percentage of single households, or the average incomes by person.

3.2 Interpretation of the characterisation at a large scale

The map visualisation of the web application depicts the normalised results by built area and helps interpret the electricity consumption at aggregated geographical levels. In the following subsections, this interpretation is discussed and compared between the residential sector and the offices, retail and public services sectors, including all the tertiary and public buildings. There are some regions without colouring because consumption

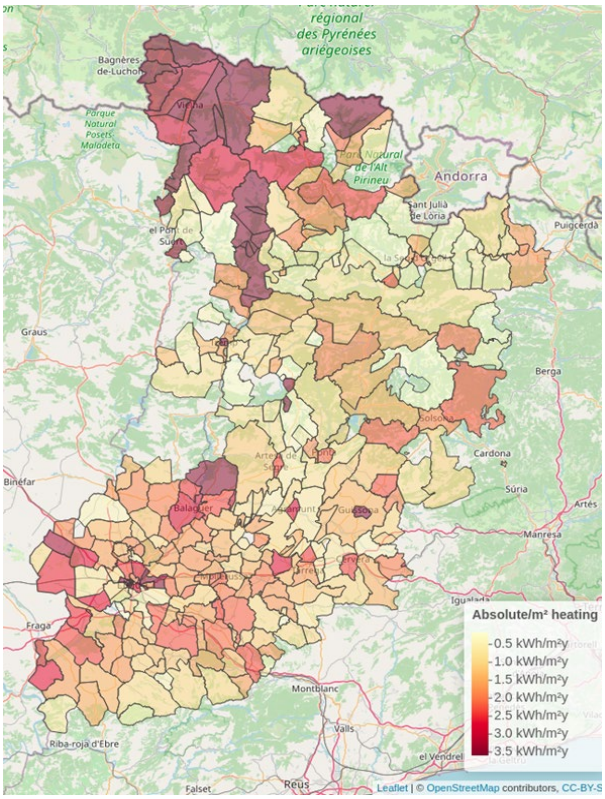
data is not available for that postal codes. Three main reasons, or a combination of them, are behind this lack of data:

1. The area does not have enough consumption points in that sector to assure privacy.
2. The implantation of the AMI is residual. It could happen in certain rural areas with few inhabitants.
3. It does not exist a sufficient amount of historical data to estimate the characterisation model parameters—at least one year of data.

3.2.1 Heating dependence

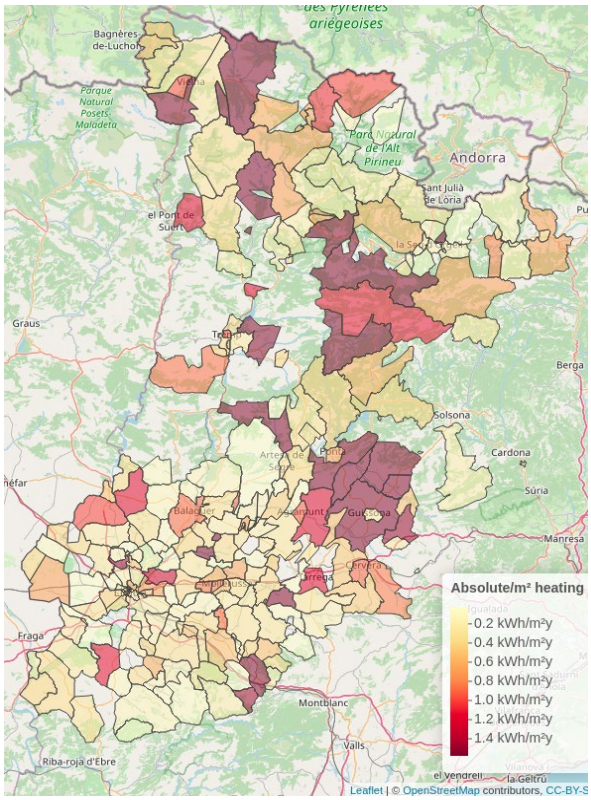
As shown in the next figures, the heating consumption depends quite highly on the location and the sector under analysis. In the case of the residential sector, there is a heating dependence between 2% to 40% of total consumption, with a special focus on certain areas of the West Pyrenees, which may be due to the lack of gas distribution in that zones and the implantation of electric radiators, heat pumps and floor heating systems. Postal codes around city centres (Lleida, Tarrega or Balaguer), where many new buildings have been constructed in recent years, also have more electricity consumption dependence due to heating. Regarding the tertiary buildings, between 3 to 30% of the total consumption is made by cause of heating uses. In this case, the dependence is not so evident in the West Pyrenees area, as many building boilers are resourced by gas-oil or biomass. Nonetheless, a significant number of postal codes, especially in the eastern part of the province, have a significant consumption due to heating.

Figure 30. Heating consumption per built area in the case of residential sector customers



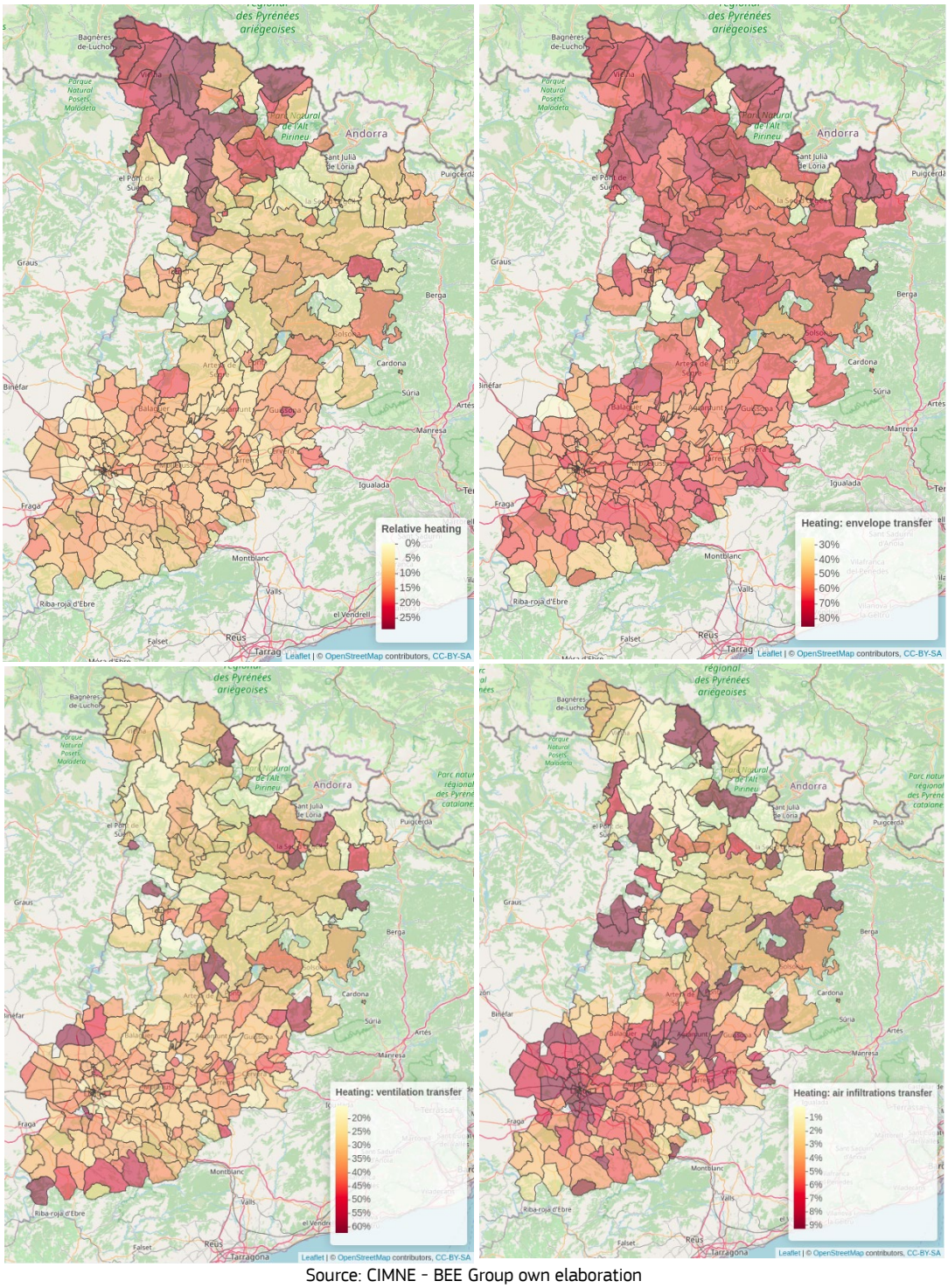
Source: CIMNE - BEE Group own elaboration and OpenStreetMap

Figure 31. Heating consumption per built area in the case of offices, retail and public services customers



Source: CIMNE - BEE Group own elaboration and OpenStreetMap

Figure 32. Heating consumption per built area in the case of residential sector customers. Estimated heat transfer coefficient factors: envelope losses, ventilation losses and air infiltration.



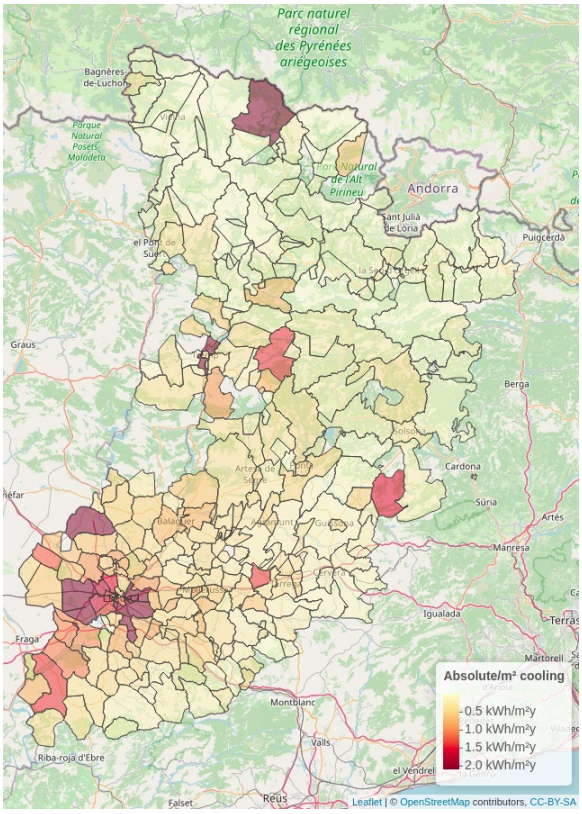
Source: CIMNE - BEE Group own elaboration

Regarding how buildings are performing and how the HTC is distributed between these three factors: envelope losses, ventilation losses and air infiltration losses, the upper figure shows that the envelope transfer tends to be the most important energy transfer factor. Additionally, it can be seen that the percentage due to envelope losses tends to increase if the postal code is located in a region where the heating dependence is higher. Regarding the ventilation and air infiltration factors, an interesting tip could be to change/repair the windows or consider more efficient user ventilation strategies, especially when the air infiltration factor becomes significant >6-7%.

3.2.2 Cooling dependence

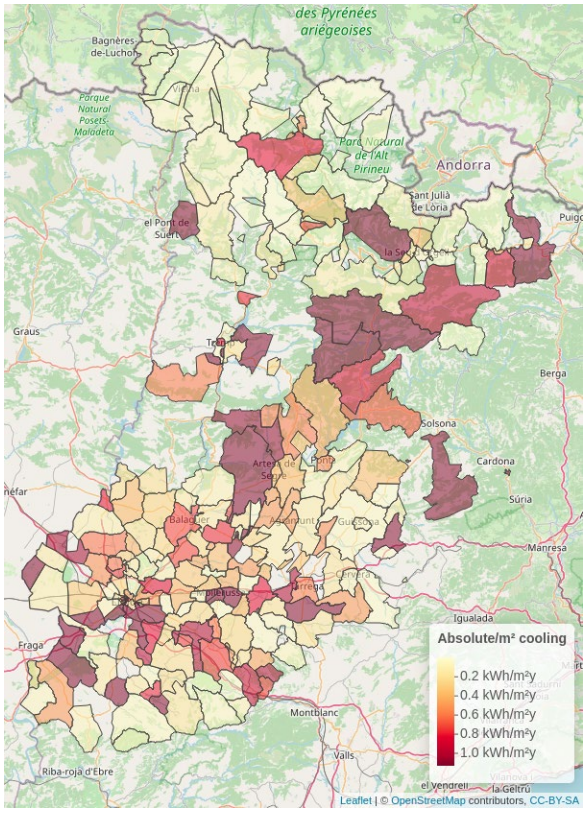
Compared to the heating dependence, the electricity consumption due to cooling needs is between 0 to 10% of the total consumption in the residential sector and between 0 to 25% in the case of tertiary buildings. As shown in the figures below, in the residential sector, the locations with more influence on cooling needs are mainly around the capital of the province (Lleida). In contrast, in the case of tertiary buildings, it is difficult to find a pattern. Still, the map could be used to indicate which regions integrate air-conditioning systems in their tertiary buildings.

Figure 33. Cooling consumption per built area in the case of residential sector customers.



Source: CIMNE - BEE Group own elaboration and OpenStreetMap

Figure 34. Cooling consumption per built area in the case of offices, retail and public services sector customers.



Source: CIMNE - BEE Group own elaboration and OpenStreetMap

3.2.3 Baseload consumption

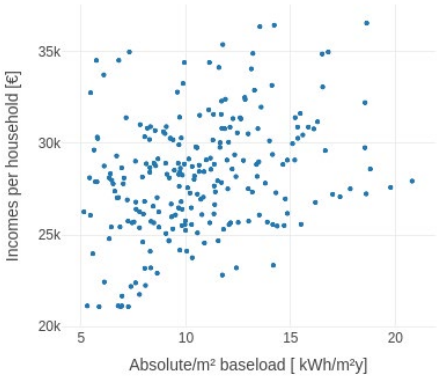
The baseload consumption has a similar geographical pattern to the cooling dependence. In addition, both sectors are the most important factor in consumption, between 55% to 97% of the total consumption, which is not strange, considering that this component aggregates all non-HVAC electric equipment (e.g. stand by devices, lighting)In the residential sector, again the regions with more baseload consumption are situated around the province's capital city. Besides, as shown in **Figure 35**, a positive correlation has been found between incomes per household and baseload consumption. This relationship is consistent because, in general, households with more financial resources have more electrical equipment and thus have a higher baseline consumption.

In the case of offices, retail and public buildings, high baseload consumption depends quite highly on the size of the buildings, as shown in **Figure 36**. The two most probable reasons are:

- smaller buildings are characterised by more electric equipment per building area compared to larger ones,
- there is a lack of energy efficiency in some smaller buildings that originates the difference in the ratio baseload consumption by building area.

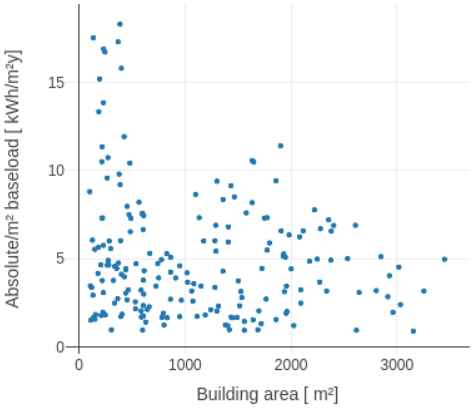
Hence, more information is needed to understand the root cause of this difference. However, it has been proved that the developed methodology provides useful insights to detect this kind of problem.

Figure 35. Correlation between baseload consumption and incomes per household in the residential sector



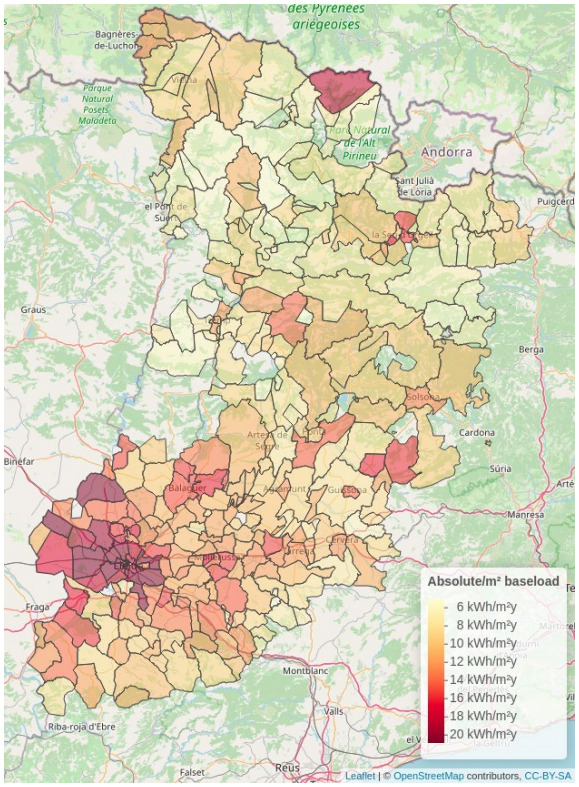
Source: CIMNE - BEE Group own elaboration

Figure 36. Correlation between baseload consumption and building size in offices, retail and public buildings



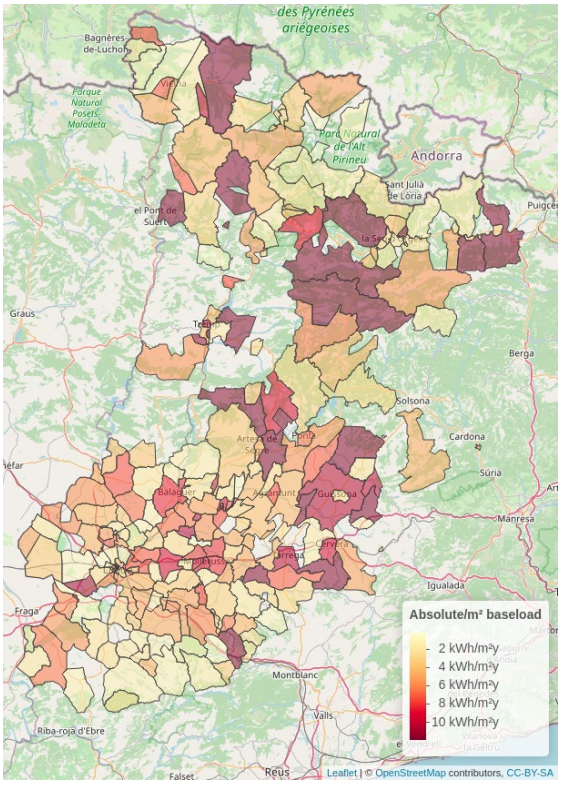
Source: CIMNE - BEE Group own elaboration

Figure 37. Baseload consumption per built area in the case of residential sector customers.



Source: CIMNE - BEE Group own elaboration and OpenStreetMap

Figure 38. Baseload consumption per built area in the case of offices, retail and public services sector customers.



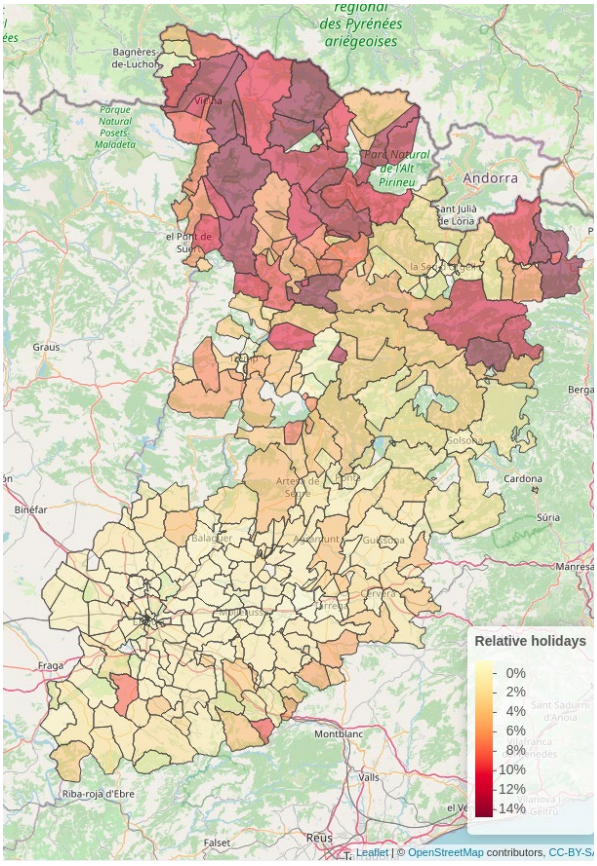
Source: CIMNE - BEE Group own elaboration and OpenStreetMap

3.2.4 Effect of holidays

In the case of the residential sector, the holidays' effect on the total consumption depends very significantly on the proximity of the postal code to the ski resorts of the Pyrenees, where lots of second residences, hotels and apartments are occupied during the weekends or holidays periods. Near the capital, the influence of holidays can even become close to negative, because many people go on short breaks or trips, whereas other people do more activities at their homes.

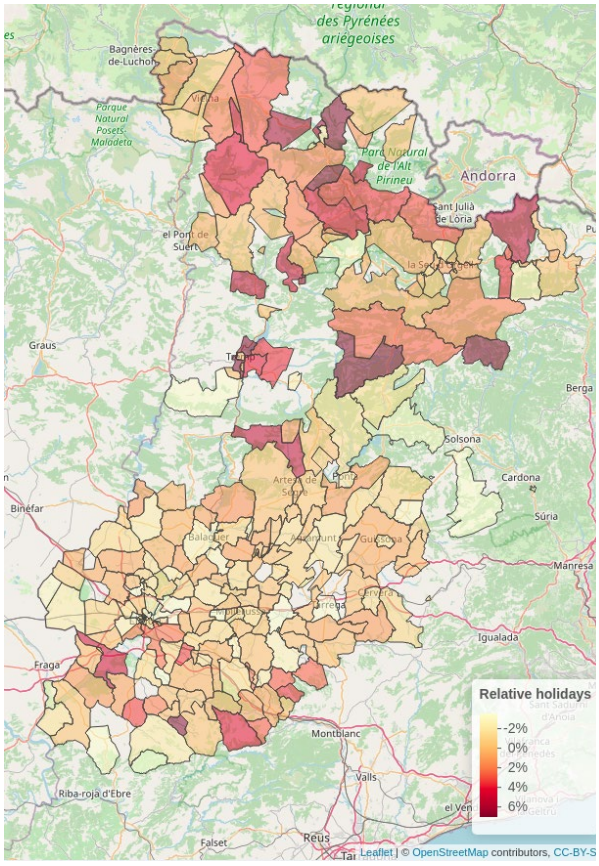
On the contrary, the influence of holidays in the tertiary sector is lower than in the residential sector. However, even if the percentages of influence range from -3% to 7%, it has the same effect in similar regions of the Pyrenees and Pre-Pyrenees, whilst, in more populated areas around the capital, the consumption is reduced compared to non-holiday periods.

Figure 39. Holidays effect on total consumption per built area in the case of residential sector customers.



Source: CIMNE - BEE Group own elaboration and OpenStreetMap

Figure 40. Holidays effect on total consumption per built area in the case of offices, retail and public services sector customers.



Source: CIMNE - BEE Group own elaboration and OpenStreetMap

4 Conclusions

A methodology to characterise different geographical areas from the electricity consumption point of view has been developed, implemented and validated.

Moreover, open-source software has been created to extract the information from publicly-available data sources, to transform and store this data into databases, to analyse and model the electricity consumption only based on high-frequency time series data and to visualise the KPIs and the outcomes obtained using a web application created on purpose.

It has been proved that the implementation of this type of methodologies is feasible in Spain, but it can also be done in other EU countries.

The list of possible applications that could reuse this characterisation methodology is quite large, targeting different types of beneficiaries:

- Public Authorities aiming at a better planning of renewables integration, prioritising the implementation of energy efficiency measures in those geographical areas where a significant fault is detected, or assessing the evaluation of ECM applied in districts or regions.
- Private companies aiming at improving their marketing strategies, based on the links existing between the territory and the electricity consumption use trends.

The same types of aggregation strategies used in this methodology could scale up/down the characterisation to wider/narrower geographical levels than postal code, such as regions, provinces, or areas affected by the same transformer station. For validation purposes, and considering that this is the lower level currently available, the Spanish postal code was used to represent what could be done with statistical models in terms of characterisation of energy consumption, merging aggregated actual electricity consumption, weather data, cadastre data and socioeconomic information.

In case of reuse of the characterisation methodology and the tools at a wider national or EU scale, it is recommended to move the granularity of the geographical levels to at least the provincial level for simplicity. However, it should be considered that the methodology and the tools, when applied at a more local level, could provide more relevant outcomes. For instance, they could be effectively used as a starting point to estimate at a local level the amount of needed in-situ renewables generation to achieve the related Sustainable Development Goals 2030. This aspect is extensively discussed in many technical departments of Spanish municipalities, and the availability in production of the tools (currently in a prototype version) would provide them accurate information about the real consumption by districts, trends and disaggregation of uses, potentiality of self-consumption. In summary, it would give them a very useful input to determine, for example, how much photovoltaic power should be integrated and where it is most needed.

References

[1] Fonseca, J.A. and Schlueter, A., 'Integrated model for characterisation of spatiotemporal building energy consumption patterns in neighborhoods and city districts', *Applied Energy*, Vol. 142, Mar. 2015, pp. 247–265.

[2] Swan, L. G. and Ugursal, V. I., 'Modeling of end-use energy consumption in the residential sector: A review of modeling techniques', *Renewable Sustainable Energy Reviews*, Vol. 13, No. 8, Oct. 2009, pp. 1819–1835.

[3] Voulis, N., Warnier, M., and Brazier, F. M. T., 'Statistical Data-Driven Regression Method for Urban Electricity Demand Modelling' *in 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe)*, Jun. 2018, pp. 1–6.

[4] 'Province of Lleida', *Wikipedia*. Oct. 25, 2020 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Province_of_Lleida&oldid=985386577

[5] 'Data Specifications / Themes / Buildings | INSPIRE' Available: https://inspire.ec.europa.eu/Themes/126/2892

[6] 'CARTOGRAFÍA CATASTRAL - INSPIRE.' Available: http://www.catastro.minhap.es/webinspire/index.html

[7] 'Experimental statistics. Main page.' Available: https://www.ine.es/en/experimental/atlas/experimental_atlas_en.htm

[8] 'DATADIS. La plataforma de datos de consumo eléctrico.' Available: https://datadis.es/

[9] 'Persona Jurídica – Sede.' Available: https://www.sede.fnmt.gob.es/en/certificados/certificado-de-representante/persona-juridica

[10] 'Codigos Postales de España.' Available: https://www.codigospostales.com/

[11] 'Productos y Servicios / Informacion estadística / Cartografía secciones censales y callejero de Censo Electoral.' Available: https://www.ine.es/ss/Satellite?L=es_ES&c=Page&cid=1259952026632&p=1259952026632&pagename=ProductosYServicios%2FPYSLayout

[12] Goeman, J., Meijer, R., Chaturvedi, N., and Lueder, M., 'Penalised: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalised Estimation in GLMs and in the Cox Model'. *R CRAN*, 2018.

List of abbreviations and definitions

AMI	Advanced Metering Infrastructure
API	Application Programming Interface
BIC	Bayesian Information Criterion
CIMNE	International Centre of Numerical Methods in Engineering (in Catalan: *Centre Internacional de Mètodes Numèrics en l'Enginyeria*)
CSV	Comma Separated Values
CVRMSE	Coefficient of Variation of the Root Mean Squared Error
DSO	Distribution System Operator
ECM	Energy Conservation Measures
EPB	Energy Performance of Buildings
ESCO	Energy Service Company
ETL	Extract, Transform and Load
EU	European Union
GIS	Geographical Information System
GML	Geography Mark-up Language
GPKG	Geo PaCKaGe
HTC	Heat Transfer Coefficient
HVAC	Heating, Ventilation and Air Conditioning
ICL	Integrated Completed Likelihood
INE	National (in Spanish: *Instituto Nacional de Estadística*)
JSON	JavaScript Object Notation
KML	Keyhole Markup Language
KPI	Key Performance Indicator
NUTS	Nomenclature of Territorial Units for Statistics (in French: *Nomenclature des Unités Territoriales Statistiques*)
MAPE	Mean Average Percentage Error
RC	Resistances and Capacitators
RMSE	Root Mean Squared Error
SQL	Structured Query Language
UML	Unified Modeling Language

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-02-21-934-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/362074

ISBN 978-92-76-40553-5