Energy Reports 7 (2021) 5667–5684

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

A data-driven method for unsupervised electricity consumption characterisation at the district level and beyond

Gerard Mor [a,*], Jordi Cipriano [a], Giacomo Martirano [b], Francesco Pignatelli [b], Chiara Lodi [b], Florencia Lazzari [a], Benedetto Grillone [a], Daniel Chemisana [c]

[a] Centre Internacional de Mètodes Numèrics a l'Enginyeria. Building Energy and Environment Group, Pere Cabrera 16 2-G, Lleida, 25001, Spain
[b] Joint Research Centre, Via E. Fermi 2749, Ispra (VA), 21027, Italy
[c] Applied Physics Section of the Environmental Science Department, University of Lleida, Jaume II 69, Lleida, 25001, Spain

ARTICLE INFO

Article history:
Received 8 June 2021
Received in revised form 7 August 2021
Accepted 30 August 2021
Available online xxxx

Keywords:
Building-stock models
Electricity
Characterisation
Data-driven

ABSTRACT

A bottom-up electricity characterisation methodology of the building stock at the local level is presented. It is based on the statistical learning analysis of aggregated energy consumption data, weather data, cadastre, and socioeconomic information. To demonstrate the validity of this methodology, the characterisation of the electricity consumption of the whole province of Lleida, located in northeast Spain, is implemented and tested. The geographical aggregation level considered is the postal code since it is the highest data resolution available through the open data sources used in the research work. The development and the experimental tests are supported by a web application environment formed by interactive user interfaces specifically developed for this purpose. The paper's novelty relies on the application of statistical data methods able to infer the main energy performance characteristics of a large number of urban districts without prior knowledge of their building characteristics and with the use of solely measured data coming from smart meters, cadastre databases and weather forecasting services. A data-driven technique disaggregates electricity consumption in multiple uses (space heating, cooling, holidays and baseload). In addition, multiple Key Performance Indicators (KPIs) are derived from this disaggregated energy uses to obtain the energy characterisation of the buildings within a specific area. 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 sector.

1. Introduction

Enhancing energy efficiency has become a priority for the European Union (Anon, 2018). Several policies and initiatives aim to improve buildings' energy performance and collect data of sufficient quality on the effect of energy efficiency policies on building stock across Europe. Knowledge about the energy characteristics of buildings and their occupants' usage is essential to define and assess strategies for energy conservation.

For the last years, dynamic measured data has been massively accessible for a significant part of the European building stock, especially electricity consumption (Anon, 2021a). Besides, accurate location-based data such as weather, cadastre and socioeconomic conditions became available with the explosion of governmental open data platforms and price-competitive weather online services. Given the recent advances in machine learning and big data processing, we are in an excellent position to develop and validate statistically-based methodologies capable of inferring, with no human interaction, the main energy features contained in the available data sets to determine how buildings perform and how their occupants consume energy at the local level. The outcomes of these data-driven methodologies can become essential to understand the building stock energy dynamics and, therefore, to support the transition to renewable and distributed generation at district or regional levels. A recent study (Fonseca and Schlueter, 2015) 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 prior unknown patterns of energy consumption and to address several ECM scenarios or, at least, to support decision-making in setting up energy transition plans. Additionally, this is the geographical scale where most of the urban transformations in Europe occur

* Corresponding author.
E-mail addresses: gmor@cimne.upc.edu (G. Mor), cipriano@cimne.upc.edu (J. Cipriano), Giacomo.MARTIRANO@ext.ec.europa.eu (G. Martirano), Francesco.PIGNATELLI@ec.europa.eu (F. Pignatelli), Chiara.LODI@ext.ec.europa.eu (C. Lodi), flazzari@cimne.upc.edu (F. Lazzari), bgrillone@cimne.upc.edu (B. Grillone), daniel.chemisana@udl.cat (D. Chemisana).

https://doi.org/10.1016/j.egyr.2021.08.195

and where the newest instruments for financing energy efficiency strategies in the building sector exist.

In literature, the energy characterisation based on modelling groups of buildings is named building stock modelling. Three major typologies of groups of buildings exist residential, industrial and services. Each of them corresponds to its own building archetypes, uses and occupancy patterns. Two main approaches for building stock modelling can be identified: top-down and bottom-up methods. Langevin et al. (2020) provided an extensive and updated literature review based on Swan and Ugursal (Swan and Ugursal, 2009) classification methods. They extended it by considering three major developments of the last ten years: big data, increased computing power, and new modelling techniques. The bottom-up approach begins with a detailed representation of a system's constituent part that is further aggregated to the whole-system level. In this case, building archetypes are used to characterise each building or a sample of buildings. The outcomes or the key performance indicators (KPIs) are scaled up to summarise the whole building stock of the analysed area. By contrast, top-down approaches begin with an aggregated view of the overall stock of the area, which is then disaggregated into subsequent sub-systems. In this approach, the energy performance of groups of buildings is analysed as a black box, in statistical terms, defined as a large sink with inputs and outputs following historical trends.

In both bottom-up and top-down approaches, energy characterisation of existing buildings at multiple geographical levels (district, city, 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. Nonetheless, it is often difficult to obtain this characterisation, which can be tackled from different viewpoints, with widely varying accuracy and associated costs. Traditionally, in the case of bottom-up approaches, the characterisation of the energy performance of a given region is performed employing Building Energy Simulation (BES) models. In these cases, a calibration of the simulated data against real monthly or annual energy consumption data should be considered since the energy performance gap between simulated and real data should be minimised. Although these models are robust, this type of calibration procedures usually ignore the changes in the behaviour of the users over time, and in many cases, the dynamics between the real consumption and the climate conditions are not properly captured. Moreover, in several methodologies, a subset of representative buildings should be considered to depict the archetype of a particular region. Therefore, this model could experience large biases against reality if the sample is not statistically significant or the calibration procedure is not properly implemented. These limitations can result in high inaccuracies in the estimates of energy performance. For the last years, data-driven techniques have been applied to bottom-up approaches to overcome the limitations of simulation-based procedures. Abbasabadi and Ashayeri (2019) presented a review paper where several data-driven techniques for urban energy modelling are classified. They detected that the future tendency should integrate data-driven models and simulation-based models, as each of them provides interesting advantages. In Voulis et al. (2018a), urban electricity demand modelling was tested for Dutch municipalities, where a combination of multiple data sets (reference electricity demand profiles, local customers composition data and aggregated local annual demand data) were used to train a regression model for local electricity demand prediction with an interesting application for local renewable energy transition plans (Voulis et al., 2018b). Kontokosta and Tull (2017) developed a predictive energy use model at the building, district, and city scales using training data from energy disclosure policies and predictors from the widely available property and zoning information. Their method was validated in New York, and the results demonstrated that electricity consumption could be reliably predicted using real data from a relatively small subset of buildings. In contrast, natural gas use presented a more complicated problem given the bimodal distribution of consumption and infrastructure availability. An interesting conclusion from this paper is that Ordinary Least Squares (OLS) methods perform better when applied to district and city scales, compared to other statistical techniques, such as Random Forest (RF) or Support Vector Machines (SVM). Oliveira Panão and Brito (2018) developed a bottom-up approach to model the aggregated hourly electricity consumption based on a Monte Carlo model. They used probability distribution functions of the building stock characteristics, web surveys for user behaviour characterisation and energy consumption data from national statistics and smart meters data sets as input of the model. The Mean Average Percentage Error (MAPE) and the Coefficient of Variation of the Root Mean Squared Error (CVRMSE) obtained during the validation of the hourly prediction against actual data are 11% and 16%, respectively. Using data from Gothenburg, Österbring et al. (2016) proposed a methodology for building-stock energy characterisation based on characteristics of the buildings, energy performance certificates, building envelope geometries from 2.5D GIS models and measured energy.

In other cases, building stock models are used as a toolbox for specific applications. For instance, in the case of Spain, a study from Romero Rodríguez et al. (2018) showed the possibilities to mitigate energy poverty in low-income districts by combining Photo-Voltaic (PV) generation and building thermal storage using actual data and calibrated deterministic models. In this case study, the authors estimated an improvement in thermal comfort of households of up to 33% in winter and 67% in summer by using individual heat pumps and the surplus production of the district PV system. Furthermore, Gouveia et al. (2019) estimated the regional energy poverty vulnerability index for Portugal at the civil parish level, based on socio-economic data, building stock characteristics, actual consumption data and theoretical consumption using the EN ISO 13790 approach.

The novelty of this paper lies in the development of a data-driven technique to characterise the electricity consumption of large areas at the district level (e.g. postal code level in Spain) and upper levels, with the particularity that actual hourly consumption is considered, which makes it quite innovative considering actual state of the art. Besides, an innovative implementation of multiple statistical techniques to model the buildings stock energy consumption is performed. It is based on inferring knowledge from actual weather data, aggregated consumption data from smart meters and building stock and socio-economic characteristics data. The aim is to obtain normalised energy trends and KPIs to describe the energy consumption of each analysed region - e.g. yearly consumption per built area or monthly-averaged daily load curve due to heating or cooling needs. This characterisation requires the implementation of modelling techniques that segment the total energy consumption into different weather-dependent and non-weather-dependent components, well-described in Section 4.

Ideally, the main final energy fuel types related to buildings should be taken into account in the building stock characterisation. The International Energy Agency (IEA), estimate that globally in 2019, and by order of importance, the main fuel types used in buildings are: electricity (32.4%), natural gas (23.4%), traditional biomass (18.5%), oil (10.5%), renewable energy (5.9%), commercial heat (4.9%) and coal (4.1%). However, multiple 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, considering the broader implementation of the Advanced Metering Infrastructure (AMI) for electricity consumption in certain EU countries, it

is much more feasible to obtain detailed sets for electricity than for the rest of them. In summary, and as a first validation of the data-driven characterisation methodology presented in this paper, electricity consumption has been considered as the only one to be characterised due to the problems in obtaining detailed data for the other main resources.

In literature, an electricity consumption segmentation at the household level using clustering techniques was developed by Kwac et al. (2014). This work helps to determine that the methodology presented in this paper need to integrate an interpreter of similar daily seasonalities, as they may not be directly related to calendar features, but to time-varying changes in the general behaviour of the consumers. In Gouveia et al. (2017) energy consumption data profiles from smart meters were used to detect active behaviour regarding space heating and cooling using the deviations from normal behaviour and survey data on socio-economic conditions, building structure, equipment and use. Even though the relatively small sample of participants (19 households with survey and smart metered data), this research enlighten the necessity to consider the non-linearity between consumption and outdoor temperature, either for cooling and heating usages. In our paper, multiple cooling and heating change-point temperatures along the day are considered as rectifiers of the model outdoor temperature regressors. The objective is to linearise their relationship, and thus, model properly their influence considering linear regression models. Furthermore, a first order low pass filter accounts for the thermal inertia of buildings, which helps to boost the model accuracy, especially when are based on data frequencies higher than daily (e.g. hourly). In more recent literature, several authors applied advanced energy signatures to model daily thermal consumption to characterise the linear and non-linear heat usage dependency on outdoor temperature, wind and solar irradiation (Rasmussen et al., 2020). Similar techniques are applied in our research, focusing on the characterisation of building stock instead of individual households. Furthermore, in Wang et al. (2021), regression and machine learning techniques were also used to detect how electricity use was influenced by weather and COVID-19 lockdowns over three large metropolitan areas city-scale aggregated forecasting (Los Angeles, Sacramento and New York). The daily models' forecasting accuracy was between 4%–6% of CVRMSE. In our paper, similar accuracy is reached 4%–12% of CVRMSE, highly depending on the number of consumers aggregated on each case. Even though, and considering the 4h-frequency aggregation considered in our analysis, the increase in error compared to the daily aggregation is very low. The results are also more accurate than the 16% CVRMSE obtained in Oliveira Panao et al. research (Oliveira Panão and Brito, 2018).

Besides the definition and implementation of the methodology, a validation case study is presented in Section 5. The outcomes are shared through a Shiny web dashboard (Chang et al., 2021) that depicts multiple plots related to the electricity consumption characterisation for each postal code and interactive maps to benchmark the whole set of KPIs, among other visualisations. The Spanish province of Lleida ($> 12\,500$ km^2) is the area selected for the case study. Section 2 extensively describe the main data sources used for the case study validation. The final goal is to provide a geographically aggregated characterisation methodology for building performance and usage trends of electricity consumption, both for the residential and public/tertiary buildings.

2. Input data

This section explains the data requirements, gathering, cleaning, and transformation procedures needed to successfully characterise the electricity consumption over the case study in Spain. Moreover, it defines the initial requirements to implement this methodology in other countries or use cases.

2.1. Cadastral data

Buildings characteristics are gathered from national cadastral datasets. The data format used by these entities across EU countries is harmonised using the INSPIRE Buildings theme (Anon, 2021b). In the case of Spain, the massive downloadable public information of cadastral datasets is available through ATOM files (Anon, 2021c), where Geography Markup Language (GML) files regarding "buildings" and "building parts" can be obtained for all the Spanish municipalities. 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, including information about 2D geometries of the building parts, gross floor area, number of floors above and below ground.
2. Typology information, including variables, such as the major current use, the total number of dwellings and building units.
3. Construction information, including the actual conditions of the building and the year of construction.

Even if the amount of information is pervasive, 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, it should be considered that many data inaccuracies can exist compared to the real conditions. Some of the encountered issues are:

- Problems dealing with buildings with several main uses (services + residential, or industrial + services), as only one use is related to each building.
- Non-realistic dwelling areas based on the gross floor area, due to the influence of large parking and/or community areas.
- Some building information is 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, up to 30% of buildings without current use information.

To avoid unrealistic estimations when aggregating this data to postal code geographical level, some filters were considered - e.g. subtract ground floors and basements from the total gross area in residential buildings with more than three floors.

2.2. 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" (Anon, 2021d). 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 income per person and household, the income primary sources, the income quantile 80 and 20 ratio, the number of inhabitants, the average population age, the percentage of people under 18 and over 65, the number of people per household, the percentage of single households, and the Gini index.

2.3. Electricity consumption data

Datadis platform (Anon, 2021e) supplies the historical hourly electricity consumption aggregated by postal code, economic sector, tariff and DSO for Spain. This platform is participated by most Spanish DSOs, who 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 (Anon, 2021f) 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 is the hourly consumption and online contracts.

In Spain, the electricity tariffs available through Datadis during the period represented in the case study (from beginning 2018 to mid-2020) are specified in Table 1.

Data within the same economic sector sometimes contains gaps, multiple energy trends, and seasonality 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 representativeness of each postal code when the results are visualised over a map.

Even considering the use of aggregated consumption data at a postal code level, which alleviates the influence of poorly measured data at some particular site, some problems were detected during the initial quality checks. Hence, it became mandatory the implementation of a data cleaning process before modelling steps. In essence, the outlier filtering avoids any measure which accomplishes, at least, one of the following conditions:

1. Hourly consumption equal to 0. It is certainly impossible to have zero consumption considering that several contracts are aggregated 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-aligned moving average plus-minus three moving standard deviations, considering a window of 15 days.

2.4. Weather data

Outdoor weather conditions are obtained through the Dark Sky API service (Apple, 2019) for the whole area in analysis. In essence, the historical weather data for the same period is downloaded for each of the postal codes considered. The most important variables in our analysis are the outdoor temperature and wind speed.

2.5. Geographical levels

Data used in the framework of this energy characterisation is related to multiple geographical levels. In this subsection, 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.

2.5.1. 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. Beyond cadastral information, and mainly due to privacy issues, there are not many other open datasets available at this level. It is worth mentioning that this geographical level would be the most interesting due to its flexibility for aggregation purposes. 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.

2.5.2. 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 (Anon, 2021g).

2.5.3. Census tract level

Census tracts are the lowest level units for disseminating statistical information and are also used to organise electoral processes. Being basically 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 most updated shapefiles of the census tract in Spain are obtained from the National Statistical Office (Anon, 2021h).

For urban areas, the census tract level offers much more detail than the postal code one. The number of building blocks inside a certain census tract is much lower than in the postal code. However, for rural areas, the representativity of both levels is very similar, as they usually represent areas of similar size.

3. The architecture of the solution

The implementation of this methodology consists of combining and analysing multiple layers of data, as shown in Fig. 1. Considering that this information has heterogeneous characteristics, both in terms of frequency, geographical reference and typology, one of the mandatory aspects regarding the 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 correlate cadastral information to socioeconomic conditions, electricity consumption and weather data. Python 3.8 (Anon, 2021i) is used to extract, transform, and load data processes, using QGIS 3.10 (Anon, 2020) as a backend to analyse geospatial data. Regarding the electricity characterisation model, it is implemented in R 4.1 (Anon, 2021j). All these scripts store the raw, intermediate and final results to a MongoDB 4 non-relational database (Anon, 2021k).

The relationships and transformations among the different databases are depicted in the UML model shown in Fig. 2, where the classes are named by the name of the provider and the name of the collection, separated using ":". In the case of intermediate or final classes used by the data analytics backend or by the frontend to visualise results, the provider's name is "beegeo". The calculations considered for the aggregations to higher geographical levels are explained following SQL format in yellow notes. The implementation of this UML representation is made using a combination of open-source analytics and

Table 1
Electricity tariffs description in the Spanish market.

Access toll name	Time-of-use structures (nº periods)	Contracted power range	Main usage
2.0	A (1) DHA (2) DHS (3)	$<$ 10 kW $<$ 10 kW $<$ 10 kW	All-kind of dwellings, houses, small-sized shops or offices
2.1	A (1) DHA (2) DHS (3)	$\geq$ 10 and $<$ 15 kW $\geq$ 10 and $<$ 15 kW $\geq$ 10 and $<$ 15 kW	Big-sized houses medium-sized shops or offices
3.0	A (3)	$\geq$ 15 kW	Public buildings, or big-sized shops, or office buildings
3.1	A (3)	$<$ 450 kW (high voltage)	Industrial buildings

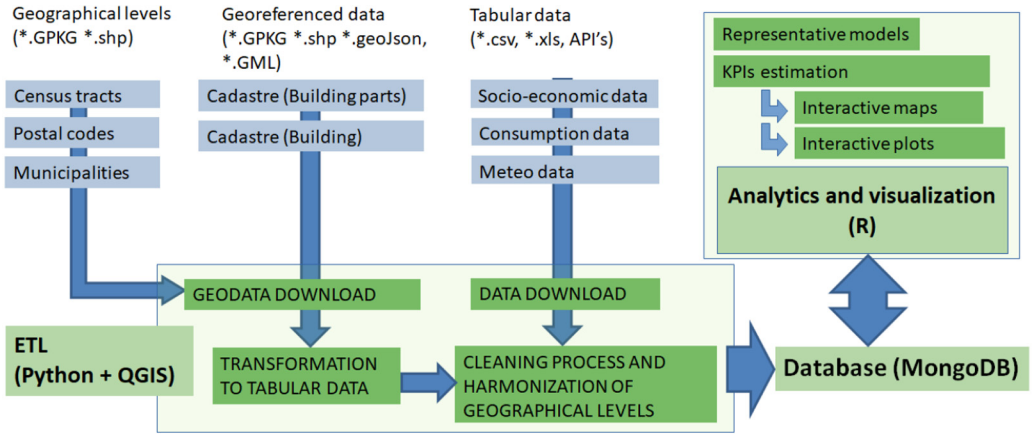

Fig. 1. General view of the data flow and the architecture of the software.

data storage technologies that allow validating the methodology over the province of Lleida. The visualisation is made using a Shiny frontend application (Chang et al., 2021), which has been developed on purpose for this case study. 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 prepared to be horizontally scalable, which is an interesting feature for future deployment of the application, either for Spain or other EU countries. The time period extends from the beginning of 2018 until June 2020, but the ETL processes are prepared to recursively obtain new data as soon as it becomes available online. To sum up, the web application is divided into four tabs: KPIs on a map, Characterisation, Benchmarking and KPIs correlation.

4. Electricity characterisation method

The characterisation methodology consists in the execution of the following steps per each region, tariff, and economic sector under analysis:

- Clustering the daily load curves to infer the most representative usage patterns.
- Estimate a regression model of the electricity consumption using calendar features, clustering results and weather conditions as exogenous variables.
- Disaggregate the raw electricity consumption in baseload, holidays, heating and cooling components.
- Calculate the performance KPIs.

4.1. Clustering model

A clustering of the daily load curves for each postal code combination, tariff and economic sector is performed to detect similar usage patterns. The representative groups obtained should be used along the algorithm to increase the reliability of the characterisation due to the consideration of the multiple seasonality's that could not be related to calendar variables 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 to obtain a model to define the typical usage patterns for each case based on the original consumption time series.

The first step is to encode the input data appropriately to the usage pattern recognition. To do so, the original hourly frequency is resampled to 4 h, as the objective is to cluster daily load curves based on their approximate peak and valley consumptions - e.g. morning consumers, double-valley consumers, or nightly consumers. Then, two normalisations and one encoding procedure are considered:

1. Conversion of the original consumption time series (Q^{abs}) to a daily relative consumption time series (Q^{rel}). $Q_t^{rel} = \frac{Q_t^{abs}}{\sum_{t \in day} Q_t^{abs}}$.
2. Generation of a matrix of days (day) as rows, and parts of the day (dh) as columns, using the daily relative consumption time series.

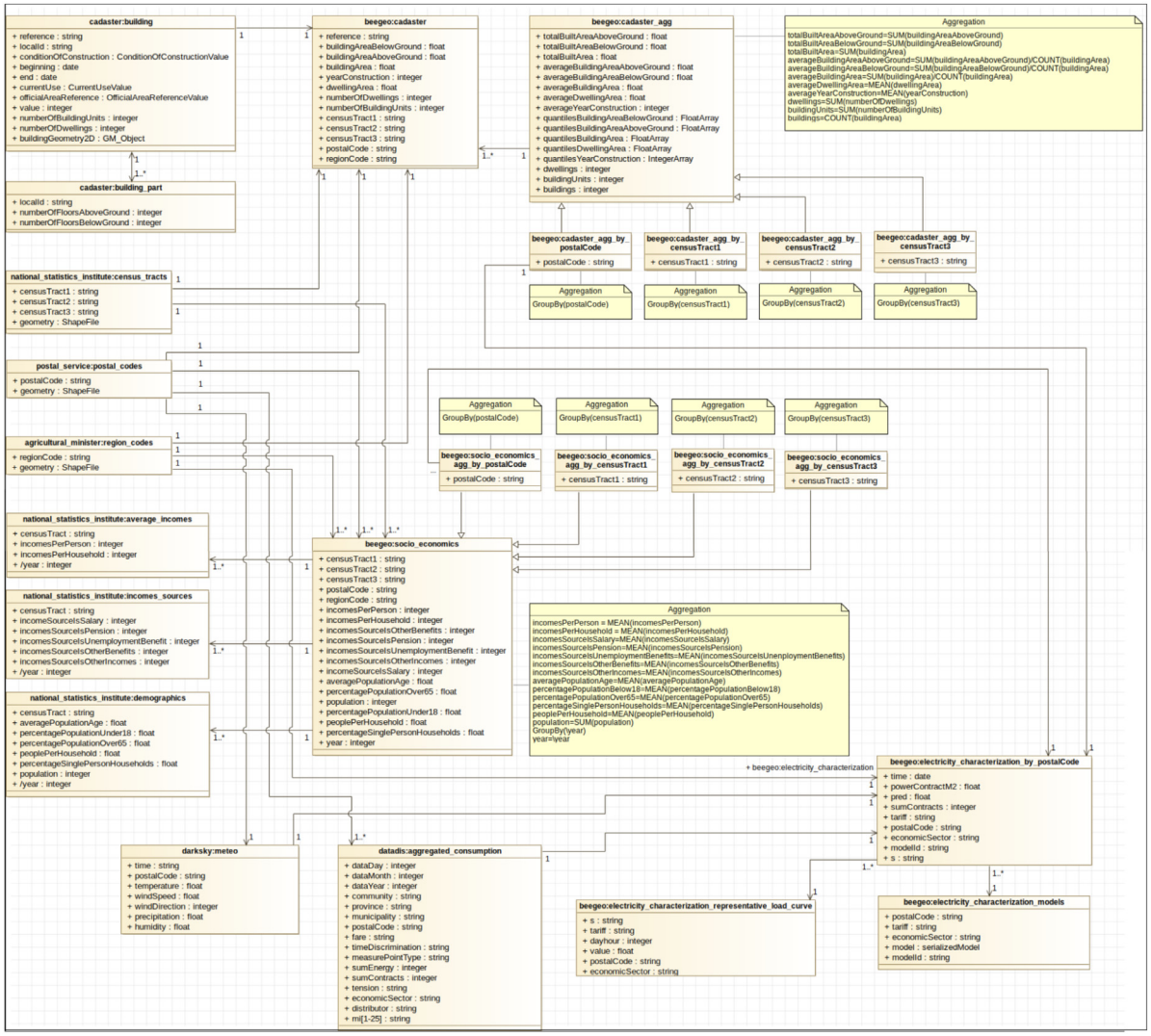

Fig. 2. UML of the used data model.

3. Transformation of the values using a Z-score normalisation, which improves the performance of the clustering algorithm.

$$Q^{z,rel}_{day,dh} = \frac{Q^{rel}_{day,dh} - mean(Q^{rel}_{dh})}{sd(Q^{rel}_{dh})}$$

Among the different clustering techniques, distribution-based clustering is 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. The distribution of every observation is specified by a probability density function through a finite mixture model of G components, as shown in Eq. (1).

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

Where $\Psi = \{\pi_1, \ldots, \pi_{G-1}, \mu_1, \ldots, \mu_G, \Sigma_1, \ldots, \Sigma_G\}$ are the parameters of the mixture model. $N_k(x_i; \mu_k, \Sigma_k)$ is the kth component Gaussian density for observation x_i with parameter vector (μ_k, Σ_k). $(\pi_1, \ldots, \pi_{G-1})$ are the mixing weights or probabilities (such that $\pi_k > 0$, $\sum \pi_k = 1$. 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 and 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 (Σ_k) structures for parameter estimation of Gaussian mixture models are the following:

- Spherical: variance is equal in all directions (where the directions are the daypart columns of the input matrix)
- 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 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.

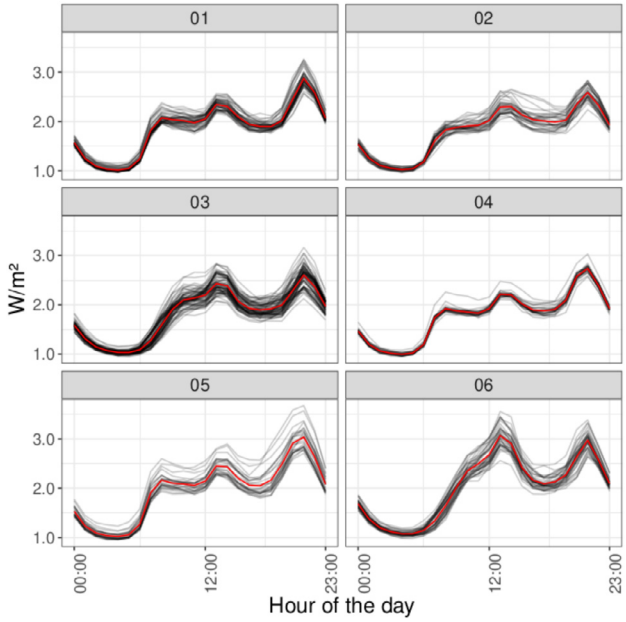

Fig. 3. Clustering of the daily load curves, only using days which are presumably not affected by weather conditions. These six profiles represent the usage patterns of the case study.

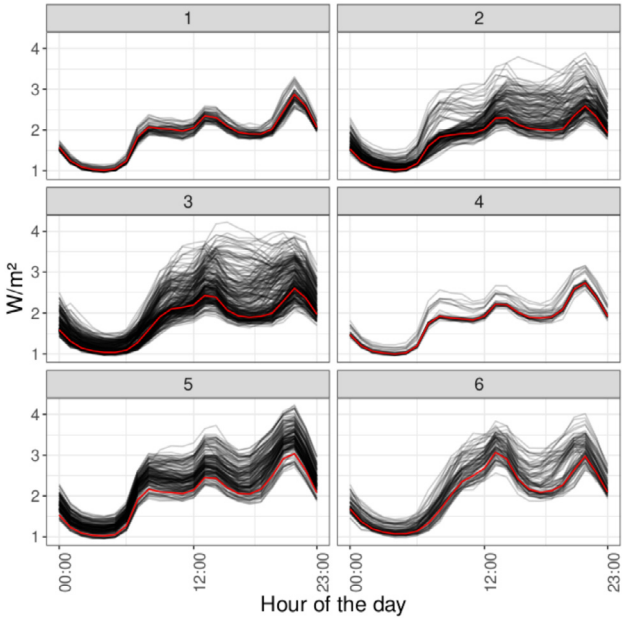

Fig. 4. Classification of the complete series using the representative usage patterns detected with the clustering technique.

Finally, an important point regarding the usage pattern detection is that to infer patterns not accounting for the weather dependence or holidays component, a clustering-classification approach with a different subset of days is considered. The clustering technique explained above 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). Subsequently, in a second step, a classification of the rest of the daily load curves is predicted using the clustering model obtained in the first stage. An example of the clustering results is depicted in Fig. 3. The red curves correspond to the usage patterns, and the black ones are the actual daily loads during the training phase of the clustering procedure. Using the same representation, the results of the classification stage are depicted in Fig. 4, where the whole period, including winter and summer seasons, are considered. As it can be seen, the weather conditions' influence tends to increase energy consumption in certain usage patterns. However, in all cases, they tend to maintain the relative shape.

4.2. Regression 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. Eq. (2) describes the major components of the penalised regression model.

$$Q_t^e = (B_t \times s_t) + (H_t \times dh_t) + (C_t \times dh_t) + \varepsilon_t \tag{2}$$

Where Q_e^t is the electricity consumption at instant t; B_t are 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 interacting with the hour of the day (dh_t). Lastly, ε_t is the error term of the model, where $\varepsilon_t \sim N(0, \sigma^2)$.

4.2.1. Baseload terms

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 the cases where input variables are periodic. The baseload terms are described in detail in Eq. (3).

$$B_t = \omega_b + S_{N_d}(p_t^d) + S_{N_w}(p_t^w) \tag{3}$$

$$S_{N_d}(p_t^d) = \sum_{n=1}^{N_d} \omega_{b,d,n,cos} \cos(2\pi n p_t^d) + \omega_{b,d,n,sin} \sin(2\pi n p_t^d) \quad p_t^d = \frac{dh_t}{24} \tag{4}$$

$$S_{N_w}(p_t^w) = \sum_{n=1}^{N_w} \omega_{b,w,n,cos} \cos(2\pi n p_t^w) + \omega_{b,w,n,sin} \sin(2\pi n p_t^w) \quad p_t^w = \frac{wh_t}{168} \tag{5}$$

Where ω_b is the linear intercept; $S_{N_d}(p_t^d)$ and $S_{N_w}(p_t^w)$ 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_t^d and p_t^w are the relative part the day or the week at instant t. The dh_t and wh_t variables mean the hour of the day and the hour of the week at instant t. The advantage of using the Fourier series is that it avoids the use 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 the clustering step). 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 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.

4.2.2. Weather dependence components

Besides the baseload terms, heating and cooling dependent components account for the consumption related to weather conditions, energy performance and characteristics of the buildings, and Heating, Ventilation, and Air Conditioning (HVAC) systems operation.

These components estimate the increase in consumption due to weather severity. They are important to 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 definition. For instance, depending on the inclusion of certain phenomena, such as ventilation or air leakage, the HTC can be different. If ventilation and air infiltration are not considered, the HTC is calculated considering the energy transfer through the building envelope, i.e. all the surrounding surfaces of the building in contact with the outdoors, ground or other buildings. If they are considered, the energy transfer due to ventilation and air infiltrations is included in the HTC definition. Furthermore, to estimate HTC some 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. Therefore, in the framework of this methodology, instead of characterising the HTC as a heat flow rate quantification, it is estimated as the change in electricity consumption, compared to the baseload, due to a variation in indoor–outdoor temperature difference. To do so, and considering that only the wind speed and the outdoor temperature are available, multiple-input transformations over these features account for the different interactions between the electricity consumption and the outdoor 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 certain situations better. 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 translated into the colder or hotter outdoor air, compared to indoor air, which enters the building, increasing HVAC systems energy consumption.

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

Finally, the third 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 retains 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, enhancing the model fitness. This transformation assumes that the dynamics of the buildings can be described by lumped parameter RC (resistance-condenser) models. In turn, this assumption means that the response in consumption due to envelope energy transfers can be modelled as a first-order low pass filter. To summarising, the space heating and cooling terms are mathematically described in Eqs. (6) and (7).

$$H_t = \omega^+_{h,lp} T^{h,lp}_t + \omega^+_h T^h_t + \omega^+_{ah} A^h_t \tag{6}$$

$$C_t = \omega^+_{c,lp} T^{c,lp}_t + \omega^+_c T^c_t + \omega^+_{ac} A^c_t \tag{7}$$

$$T^{h,lp}_t = (T^{bal,c}_{dh_t} - T^{o,lp}_t) d_{s_t} \qquad T^{c,lp}_t = (T^{o,lp}_t - T^{bal,c}_{dh_t}) d_{s_t}$$

$$T^h_t = (T^{bal,h}_{dh_t} - T^o_t) d_{s_t} \qquad T^c_t = (T^o_t - T^{bal,c}_{dh_t}) d_{s_t}$$

$$A^h_t = W^s_t T^h_t d_{s_t} \qquad A^c_t = W^s_t T^c_t d_{s_t}$$

$$T^{o,lp}_t = \begin{cases} \alpha T^o_t & \text{if } t = 0, \\ \alpha T^o_t + (1-\alpha) T^{o,lp}_{t-1} & \text{if } t > 0. \end{cases} \qquad \alpha = 1 - e^{-t_{sampling}/(2\pi\tau/24)}$$

$$d_{s_t} = \begin{cases} 1 & \text{if weather dependence in } s_t, \\ 0 & \text{if no weather dependence in } s_t. \end{cases}$$

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 related to the heat losses through the envelope and is calculated as the difference between balance heating temperature ($T^{bal,h}_{dh_t}$) at the portion 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 ($T^{bal,h}_{dh_t}$) at the part of the day (dh_t) and the raw outdoor temperature (T^o_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); $\omega^+_{c,lp}$ 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 related to the heat gains through the envelope and is calculated as the absolute difference between balance cooling temperature ($T^{bal,c}_{dh_t}$) at the part of the day (dh_t) and the low-pass filtered outdoor temperature ($T^{o,lp}_t$); ω^+_c is the always-positive linear coefficient for the raw cooling

dependent term(T_t^c), which is usually related to ventilation heat gains, and it is calculated as the difference between balance cooling temperature ($T_{dh_t}^{bal,c}$) at the part of the day (dh_t) and the raw outdoor temperature (T_t^o); ω_{ac}^+ is the always-positive linear coefficient for the heat gains due to air infiltrations (A_t^c), which is the wind speed (W_t^s) multiplied by the raw cooling dependent term (T_t^c). Besides, the α value of the low-pass-filtered outdoor temperature depends on the $t_sampling$, which is the number of measures per hour of consumption time series Q^e, and the τ thermal time constant, which defines the number of hours that the synthetic building reacts over a certain change in outdoor temperature. 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 (ds_t).

4.2.3. Impact of holiday seasonality

After the first tests of the implementation, the authors detected that the influence of holidays tends to generate significant change points in electricity consumption for certain regions, sectors and periods of the year. In most cases, the holidays periods occurred in correspondence of national holidays, Fridays or Mondays between national holidays and weekends, winter and summer weekends, and the summer vacations. However, it was difficult to find a feature that linearly correlates the holidays component of the electricity consumption with the different local festivities of every region along the year. As a first attempt, some of the features that could be used are the number of tourists, second homes occupancy, or hotel bookings at the postal code level and daily frequency. However, 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 and the residuals between the actual and predicted data during the holidays period are considered as the holiday's component. In addition, this holidays dependence component is estimated only when a difference of at least 20% is detected between the RMSE of the holidays/non-holidays period.

4.2.4. Impact of COVID-19 lockdown periods

The Covid-19 Spanish lockdown, during the period from March 15th to June 21st 2020, significantly affected the energy consumption either in residential, industrial or public sectors. Changes in business activities, user behaviour and building occupancy caused this situation. For the presented case study, the time period analysed depends on the availability of electricity consumption data for each postcode. In general, the evaluated period comprised mid-2018 to mid-2020. Thus, the data used to validate the characterisation methodology was fully affected by this lockdown period. A set of terms have been introduced into the regression model to quantify the decrease or increase in consumption due to the lockdown. They basically add 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 this period is that holidays effect on energy consumption must be fixed to zero, as people should have stayed at home for those periods, except in particular cases.

4.2.5. Training of the model

The electricity time series considered during the training phase changes slightly depending on the economic sector considered. It clearly depends on the most representative area factor for each economic sector, as the characterisation outcomes are further compared among different regions. The built area normalisation becomes a key factor in assessing 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, as is shown in Fig. 5. This procedure provides information on how the reference building is evolving in time. So, the characterisation coefficients become, in some sense, time-variant. To decrease the computational time, the original hourly frequency of the input time series is resampled to 4 h.

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 (Goeman et al., 2018), 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. The optimisation of these coefficients is made using a Genetic Algorithm (GA) that iterates and evolves chromosomes (in this case are the binary representation of the parameter values to optimise), minimising a cost function, which in this case is the Root Mean Square Error (RMSE) of the predicted consumption versus the metered consumption data. As a required initial input for the GA, a range of feasible values for each parameter to estimate is defined. In the case of $T_d^{bal,h}h$, the heating balance temperature range goes from 10 to 22 °C, in steps of 0.5. For $T_d^{bal,c}h$, the cooling balance temperature ranges between 18 to 30 °C, in steps of 0.5. The building thermal inertia parameter (τ) ranges between 1 to 48 h in steps of 1. Finally, the boolean activators for the weather dependence in each daily seasonality (d_s) can be 0 or 1. In each training period, the initial parameters considered for the GA optimisation are the ones obtained in the last period, that is the reason to increase the number of maximum iteration permitted in the case of the first training period (50 vs. 20), when no initial values are available. The population considered in the GA is 300 for each iteration and the elitism in set to a 5%.

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

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

Unknown terms for each day part: $\omega_{h,lp}^+$, ω_h^+, ω_{ah}^+, $\omega_{c,lp}^+$, ω_c^+, ω_{ac}^+, $T_{dh}^{bal,h}(^*)$ and $T_{dh}^{bal,c}(^*)$.

(*) Estimated using a genetic algorithm optimiser

5. Case study results

Rather than summarise in detail the results over the whole province of Lleida (Spain), which might be investigated in future studies, consumers in the residential sector of postal code 25006 are selected to show the intermediate and final results obtained

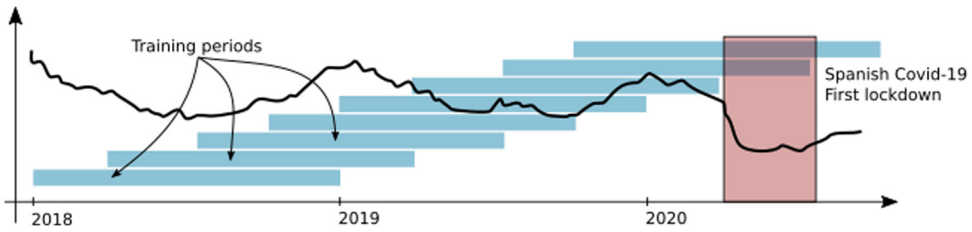

Fig. 5. Model training periods to characterise the evolution in time of the dependencies.

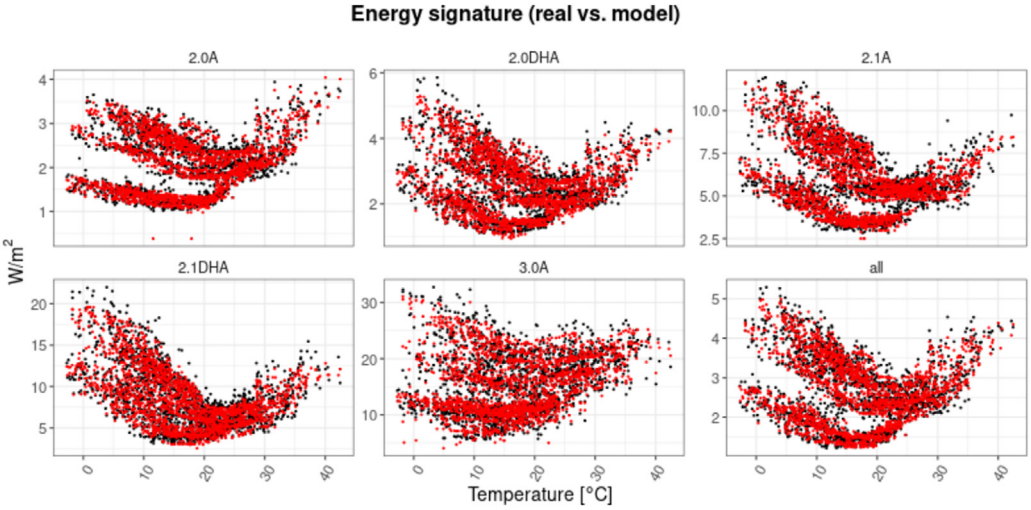

Fig. 6. Predicted 4-hourly aggregated energy signature versus actual data.

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

Period–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

Table 3
Coefficient of Variation of the Root Mean Squared Error (CVRMSE) over distinct periods and tariffs.

Period–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

during the validation procedure. This helps to focus on each of the results obtained concerning the models' accuracy and the estimated KPIs linked to the energy performance of buildings and usage patterns of their occupants.

5.1. Characterisation of a postal code

The postal code analysed is related to the Zona Alta neighbourhood in the city of Lleida. It is known as one of the most well-being districts in Lleida, at least compared to those near the city centre. Some of its socio-economic characteristics are household incomes of 36,498€ per year, incomes quantile 80–20 ratio of 3.23 (one of the highest of the province, which means there are large differences between low and high salaries), an average population age of 47.42 years, with 26.95% of people older than 65 and 13.59% under 18.

5.1.1. Estimated model

The accuracy of the models for each of the tariffs and evaluation periods are detailed in Tables 2 and 3. For both of the selected metrics, the average accuracy (MAPE: 5,04%, CVRMSE: 6,51%) is very high considering the characterisation purposes of this methodology. Even dealing with 4h-frequency predictions, the accuracy level reaches the state-of-the-art forecasting techniques at the city-scale level and daily aggregation. Fig. 6 shows the energy signature between the 4h-resampled real observations and the predictions of the models. It has been proved that the predictions capture the main trend of the original data, and even the variance is extremely similar.

The weather-related coefficients are depicted in Fig. 7. Dark blue lines correspond to the characterisation coefficients between June 2019 to May 2020 and the yellow ones from July 2018 to 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 on the dh_t, τ is the thermal time constant of the building, U_{raw} cooling values are the ω_c^+ model coefficients depending on dh_t, U_lp cooling values are the $\omega_{c,lp}^+$ model coefficients depending on the dh_t, I^{air} cooling values are the ω_{ac}^+ model coefficients depending on the dh_t, and T^{bal} cooling values are the cooling balance temperature depending on the dh_t. It can be seen that the coefficients across different tariffs vary largely and tend to be higher the more electricity is consumed by

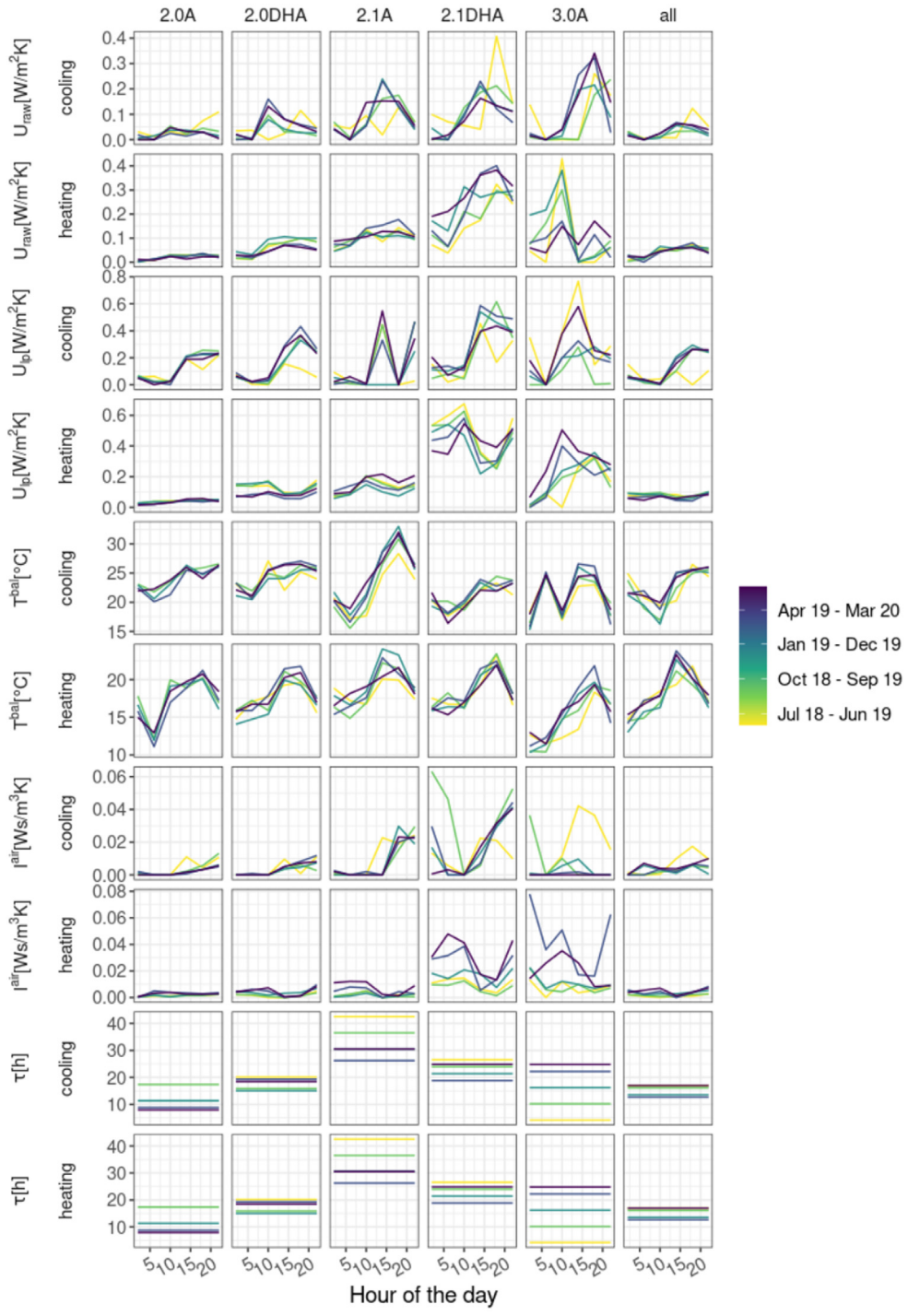

Fig. 7. Weather-dependent characterisation parameters of the model.

the tariff customers. This is a normal effect, as customers with 2.1 and 3.0 tariffs tend to have more domestic appliances or electrical driven HVAC equipment in their households. One of the most interesting insights is that space heating and cooling dependencies tend to differ widely along day time, responding with more emphasis to weather conditions during sunlight hours. Moreover, the estimated balance temperature helps to understand the most common HVAC operation schedule during a typical day, or, in other words, how people or energy managers tend to set the thermostats. Additionally, differences in the thermal time constant show variations in building's envelope characteristics between tariffs. At first glance, it seems that the 2.1 A tariff is more related to higher thermal inertia buildings, which could also be related to better-insulated buildings. Regarding the baseload characterisation, each usage pattern's daily and weekly profile and tariffs are obtained using the model parameters.

In summary, using the developed regression model, the decomposition of the three main components of buildings electricity loads (baseload, space heating and cooling) is made for the whole period of data within each of the evaluation periods (from June 2018 to May 2019, and from June 2019 to 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, Fig. 8 represents the daily disaggregation and

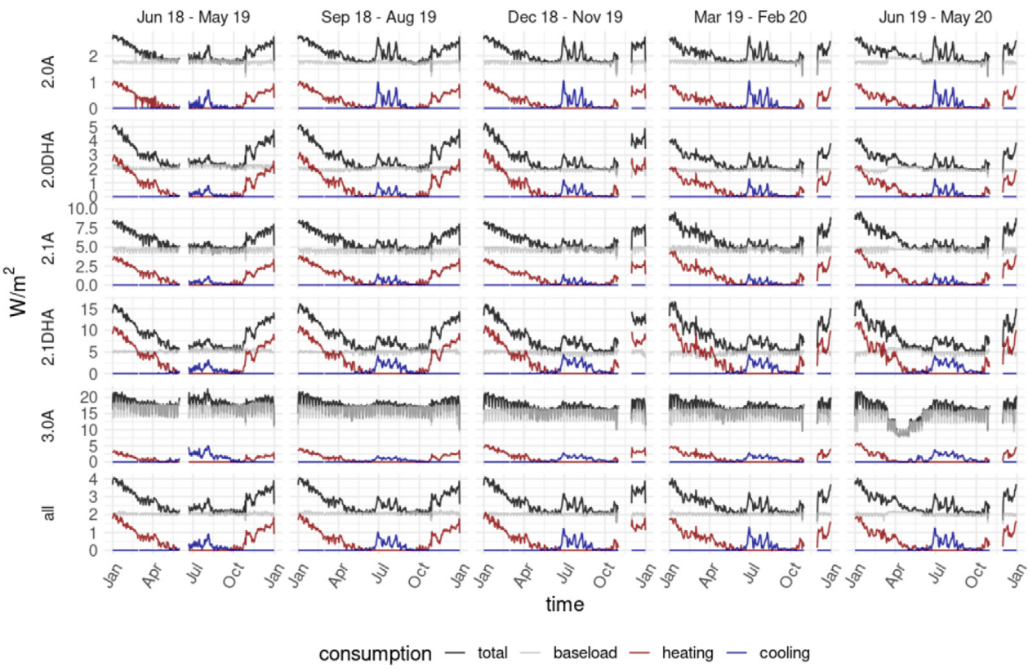

Fig. 8. Daily electricity disaggregation results over distinct periods and tariffs.

the total consumption. To compare the yearly evolution between different periods, the *X*-axis represents the months from January to December.

From Fig. 8, it can be noted that, in all the cases, the significance of the baseload consumption is much higher than the weather dependence components. Also, the high variance in the baseload component in tariff 3.0 A 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 3.0 A, where allegedly some business buildings/dwellings are integrated into the residential sector subset of the Datadis database. The evolution of the heating and cooling components through the year seems to fulfil the expected 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 a 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.

5.1.2. Summarised KPIs

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

The results of the clustering and classification of the usage patterns are illustrated in Fig. 9. In the right pane, the different usage patterns in multiple colours are depicted, and 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.

The heat map shown in 10uses 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 confinement, it has been incremented by about 20% during the daytime period from 12 h to 16 h. This can be related to more people in their homes interacting with electricity-driven cooking systems during lunchtime. In contrast, 3.0 A customers decrease their consumption drastically during those months. Regarding the heating and cooling components, it can be observed that the different intraday dependencies along different tariffs and months of the year (see Fig. 10). Maybe, again, the 3.0 A customers clearly behaved significantly different in terms of cooling dependency compared to customers with other tariffs. Besides increasing the understandability of the distribution between components and their evolution in time, Fig. 11 represents the relative disaggregation, on a natural year basis, between the baseload, the heating and the cooling components, and the impact of holidays and Covid-19 lockdown on the total consumption.

For instance, concerning tariff 2.0 A 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 do not contribute at all. In this case, the Covid-19 lockdown had a shallow impact during the lockdown period (March 15th to June 21st 2020). Another conclusion is that the evolution of the different components in time is rather similar. However, large differences can be detected between different tariffs, and this corresponds to the different users/building typologies that characterise each tariff.

Besides the relative disaggregation, the web application also provides the point of view of the absolute consumption contribution in kWh per natural year. Using this representation, a decrease in total consumption for tariffs 3.0 A and 2.0DHA is detected, especially the former, which is much more affected by the Covid-19 lockdown (approx. -20%, according to the relative segmentation results). Then, in general, for the rest of the tariffs, the same amount of total consumption during the whole evaluation period is observed.

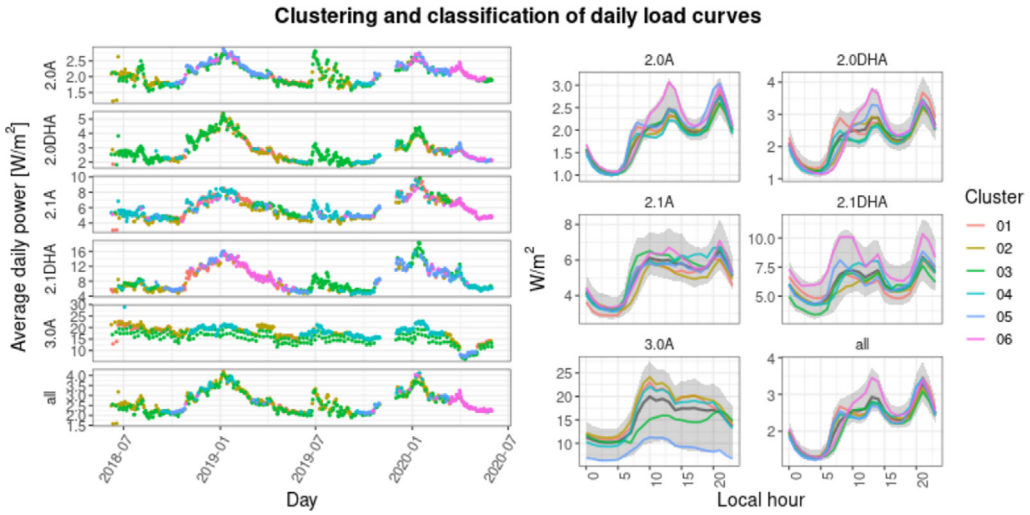

Fig. 9. Usage patterns detected over distinct tariffs.

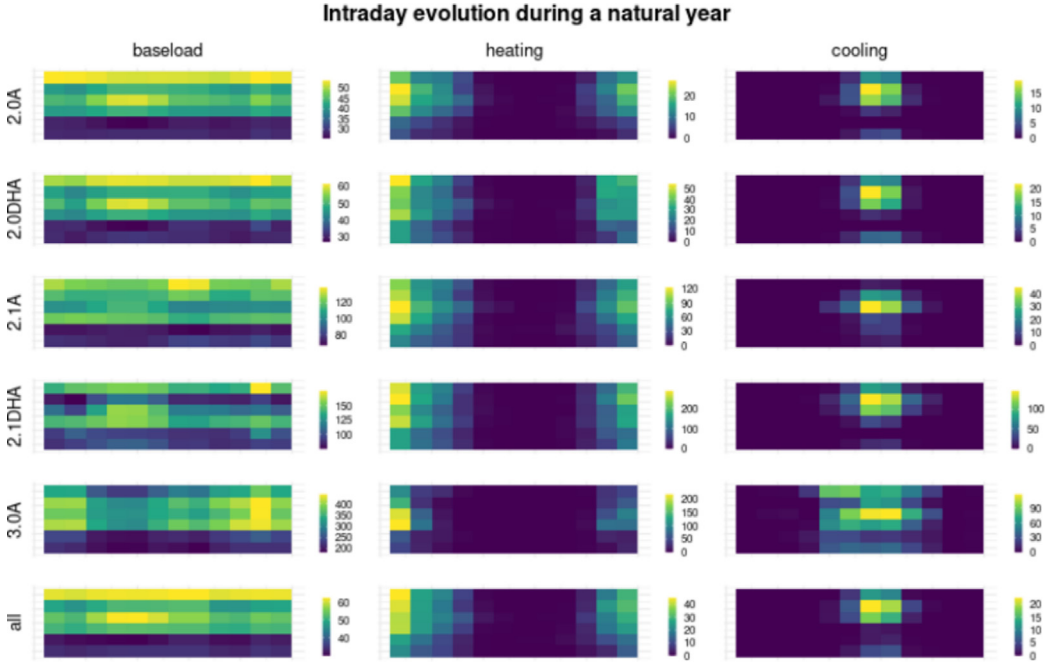

Fig. 10. Intraday summarised electricity disaggregation results over a natural year and distinct tariffs.

5.2. Results at a province level

The characterisation results over the whole province will be described in further research publications. However, to show the web dashboard created for this purpose, a set of examples are described in the following paragraphs. This validation has been launched on a single server equipped with a 12-core 3.6 GHz CPU and 32 GB RAM. The execution of the model training algorithm and the calculation of all the KPIs related to all the historical periods available and all combinations of economic sector, postal codes, and tariffs available within the province of Lleida, took 18h. Once the aggregated consumption dataset of the whole month is gathered, the analysis can be reassessed, considering the new data, in less than 2.5 h. It means that the batch calculation on the same conditions for all the Spanish provinces would take less than six days. This computational cost is totally affordable considering the low cost of this type of server and a monthly basis update of the characterisation.

Fig. 12 depicts the home section of the dashboard, whose purpose is to give a clear and simple visualisation of all the estimated consumption KPIs, cadastre information and socio-economic indicators on a map. The visualisation can be filtered by tariffs, economic sectors, periods, percentiles ranges. An interesting feature is a tiny histogram representing the distribution of values of the variable depicted on the map, especially when outliers can generate useless colouring legends.

The characterisation tab, shown in Fig. 13, represents the complete assessment of the electricity consumption of a specific postal code and economic sector selected over the map. Several of the plots shown in this tab are interactive versions of the

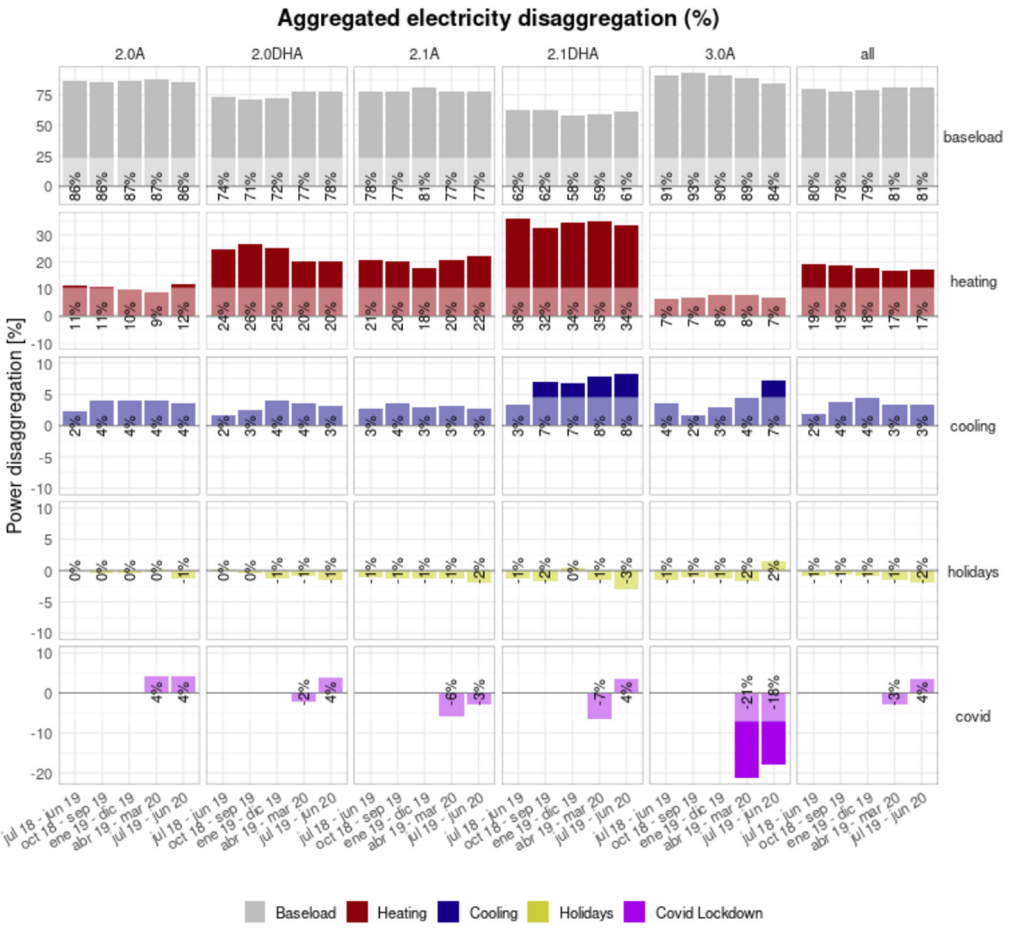

Fig. 11. Yearly-aggregated relative segmentation of the electricity consumption over distinct periods and tariffs.

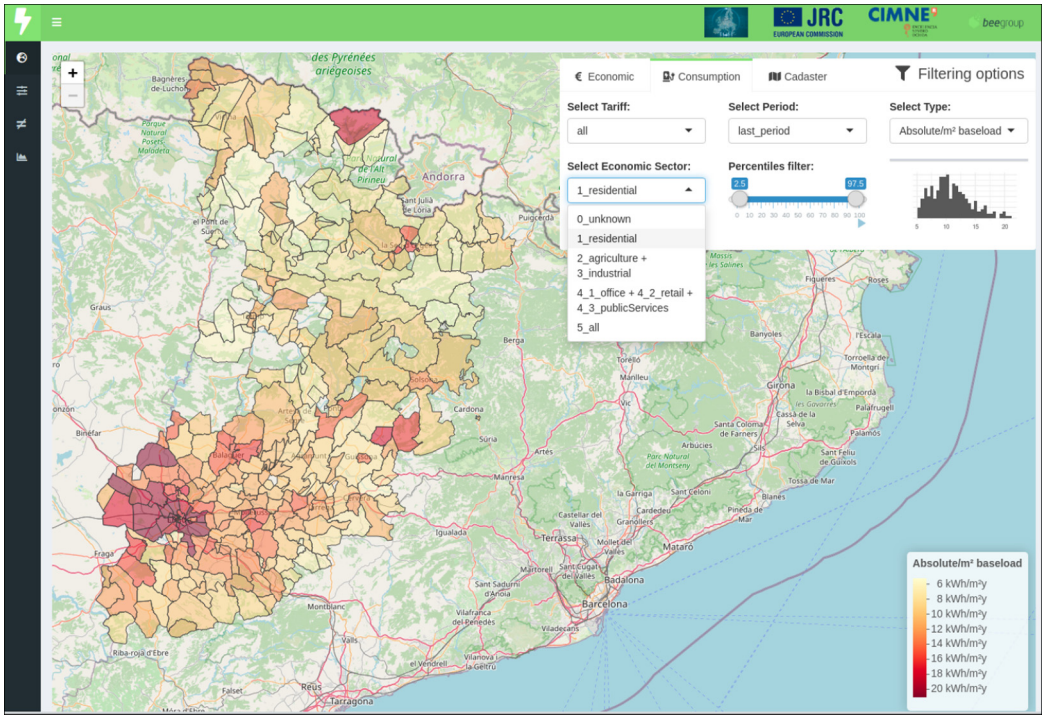

Fig. 12. Web application - "KPIs on a map" tab.

summarised KPIs explained in the subsection above, such as information about the model accuracy, the usage patterns detected and the disaggregation results in several time aggregations. The user can go deeper into the most common electricity uses over a certain geographical area.

In Fig. 14, the benchmarking tab is depicted, where the objective is to exploit the usage of the characterisation models to compare in detail two postal codes. This comparison is made by the estimated electricity components, normalising the results of the second postal code to the weather conditions and building/dwelling sizes of the first one. This normalisation procedure means that the divergence in electricity consumption should be caused by the difference in the energy performance of buildings, alternative usage patterns in electric devices, or by a different HVAC systems operation in cooling and heating electricity consumption components. In parallel, intraday differences along a natural year between the baseload consumption, and the impact of holidays and the Covid-19 lockdown period, are also represented.

Finally, Fig. 15 shows the tab that allows cross-correlating all the KPIs to understand tendencies and relations between them, providing a wider interpretation of the territory and understanding if the variation of a certain cadastre or socio-economic indicator has a significant correlation to another estimated energy consumption KPI. For instance, it could be inferred if there is a relation between holidays periods contribution to the energy consumption and average percentage of single households, or the average annual incomes per person.

6. Conclusions

A methodology to characterise actual electricity consumption of large geographical areas has been developed, implemented and validated. It has been proven that the segmentation of the aggregated electricity time series provides multiple interesting possibilities to estimate KPIs related to energy performance buildings and occupants usage trends. Moreover, it has been developed an open-source platform able to extract information from publicly available data sources. This platform is split into two main parts: a back-end and a front-end. The former gathers, transforms and stores the data into databases. These data are accessible to data analysis tools designed to model the buildings' electricity consumption only using high-frequency time series data of actual consumption and weather data as the main inputs. The latter visualises the KPIs and the obtained outcomes through a purpose-built web application. This research demonstrated that implementing this type of data-driven methodologies is feasible for large regions in Spain. Still, other European countries can also apply it as long as similar open data sources are available. The list of possible applications that could use the methodology and the web platform is pretty extensive, targeting different types of beneficiaries:

1. Public authorities interested in improving the understanding of the energy consumption flows within their territory, producing better planning and optimal integration of renewable energies, prioritising the ECM implementation at the local level, or assessing ECM impacts over districts or regions.
2. Private companies aiming at improving their marketing strategies based on the existing links between the territory and the electricity consumption use trends.

Finally, 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 global energy data-driven characterisation

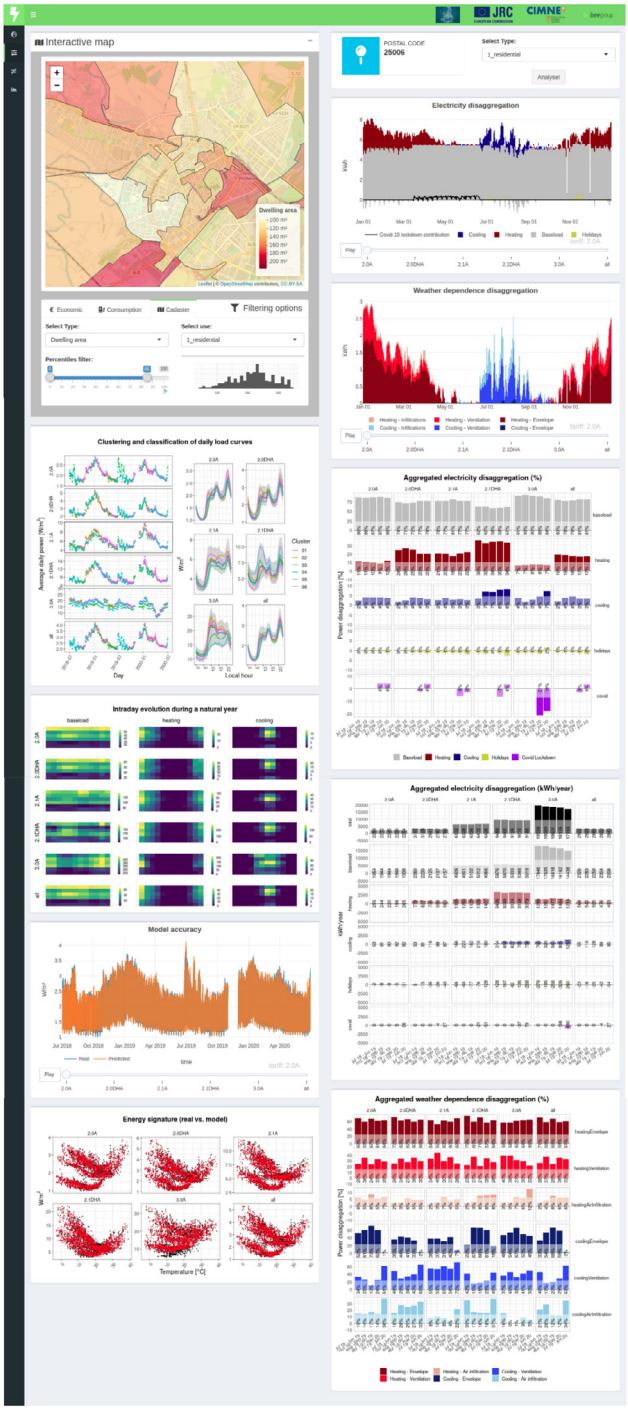

Fig. 13. Web application – "Characterisation" tab.

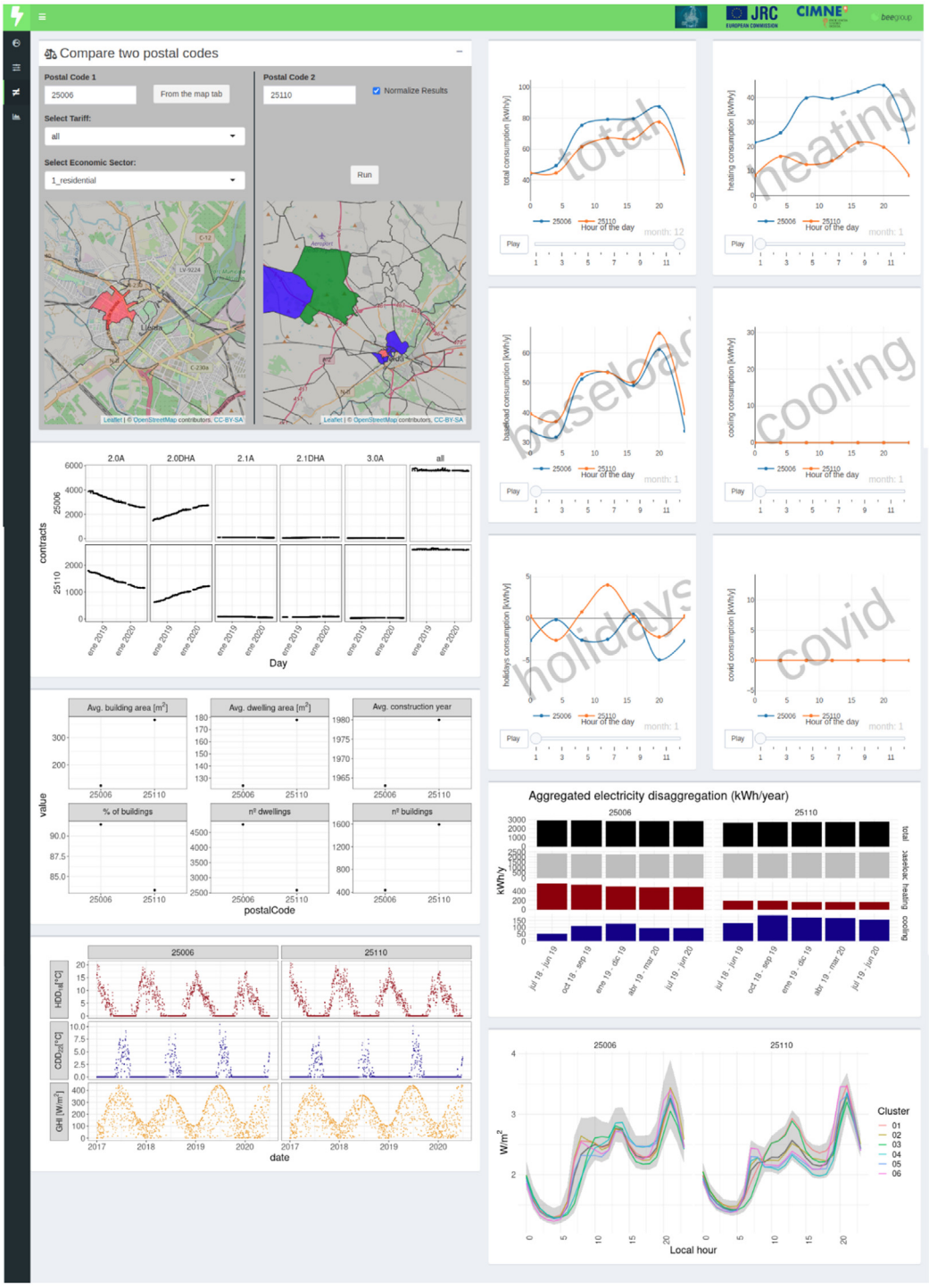

Fig. 14. Web application – "Benchmarking" tab.

techniques due to the pronounced tendency to electrify all-kind of building systems and the strong implementation of advanced meters for gas consumption.

To sum up, practical applications that could use the outcomes presented in this characterisation, have to assume that the methodology was only tested with electricity consumption. The inclusion of other final energy fuel types should slightly vary the data-driven modelling approach presented in this paper and would require another validation procedure with actual data.

CRediT authorship contribution statement

Gerard Mor: Conceptualisation, Methodology, Formal analysis, Software. **Jordi Cipriano:** Writing – original draft, Methodology, Writing – review & editing. **Giacomo Martirano:** Project administration, Methodology, Writing – review & editing. **Francesco Pignatelli:** Supervision, Project administration, Funding acquisition. **Chiara Lodi:** Writing – original draft, Writing – review & editing. **Florencia Lazzari:** Data Curation, Visualisation. **Benedetto Grillone:** Data Curation, Visualisation. **Daniel Chemisana:** Supervision.

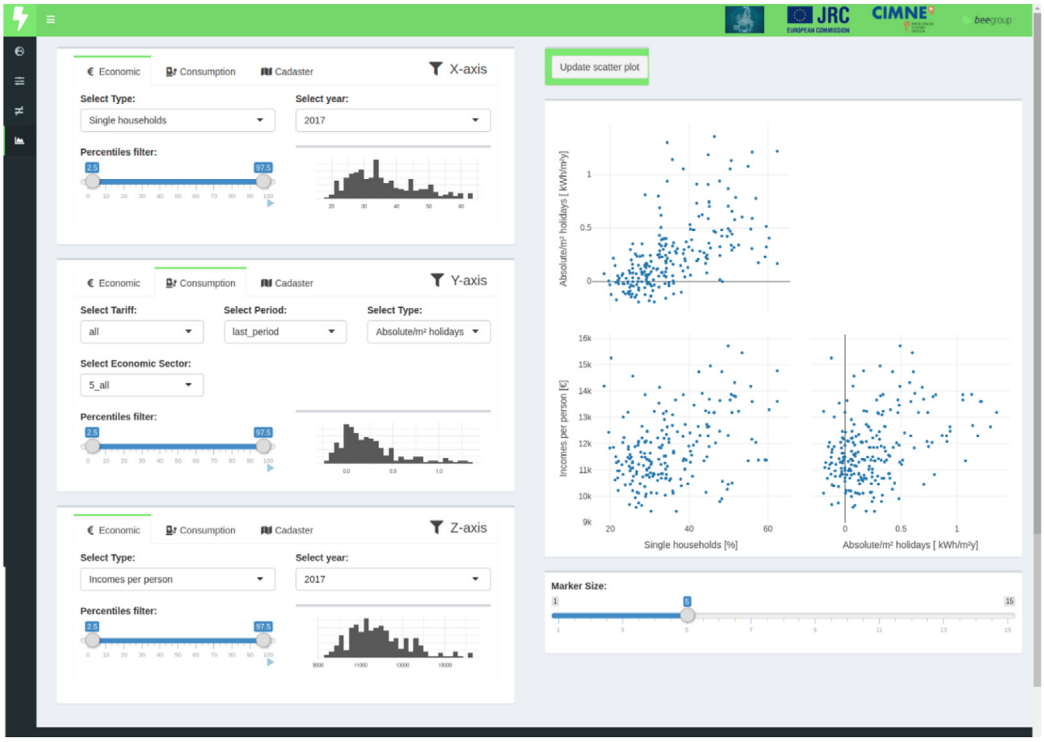

Fig. 15. Web application - "KPIs correlation" tab.

Declaration of competing interest

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

Acknowledgements

This work emanated from research conducted with the financial support of the European Commission through the H2020 project BIGG , grant agreement 957047, and the JRC Expert Contract CT-EX2017D306558-102. D. Chemisana thanks ICREA for the ICREA Acadèmia. Dr J. Cipriano also thanks the Ministerio de Ciencia e Innovación of the Spanish Government for the Juan de la Cierva Incorporación grant.

References

Abbasabadi, Narjes, Ashayeri, Mehdi, 2019. Urban energy use modeling methods and tools: A review and an outlook. Build. Environ. 161, 106270.

Anon, 2018. Directive (EU) 2018/844 of the European parliament and of the Council of 30 may 2018 amending directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency (Text with EEA relevance). OJ L, Code Number: 156.

Anon, 2020. Welcome to the QGIS project. https://www.qgis.org/en/site/.

Anon, 2021a. Smart metering deployment in the european union | JRC smart electricity systems and interoperability. https://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union.

Anon, 2021b. INSPIRE data specification on buildings – technical guidelines. https://inspire.ec.europa.eu/id/document/tg/bu.

Anon, 2021c. Cartografía catastral. http://www.catastro.minhap.es/webinspire/index.html.

Anon, 2021d. Instituto Nacional de Estadística - Estadística experimental. https://www.ine.es/en/experimental/atlas/experimental_atlas_en.htm.

Anon, 2021e. DATADIS. La plataforma de datos de consumo eléctrico. https://datadis.es.

Anon, 2021f. Fábrica Nacional de la Moneda y Timbre. https://www.sede.fnmt.gob.es/en/certificados/certificado-de-representante/persona-juridica.

Anon, 2021g. Codigos Postales de España. https://www.codigospostales.com.

Anon, 2021h. Cartografía secciones censales y callejero de Censo Electoral. https://www.ine.es/prodyser/callejero/.

Anon, 2021i. The official home of the Python Programming Language. https://www.python.org/.

Anon, 2021j. The R project for statistical computing. https://www.r-project.org/.

Anon, 2021k. The most popular database for modern apps. MongoDB. https://www.mongodb.com.

Apple, 2019. Dark sky API. https://darksky.net/dev.

Chang, Winston, Cheng, Joe, Allaire, J.J., Sievert, Carson, Schloerke, Barret, Xie, Yihui, Allen, Jeff, McPherson, Jonathan, Dipert, Alan, Borges, Barbara, 2021. shiny: Web application framework for R.

Fonseca, Jimeno A., Schlueter, Arno, 2015. Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. Appl. Energy 142, 247–265.

Goeman, Jelle, Meijer, Rosa, Chaturvedi, Nimisha, Lueder, Matthew, 2018. penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the cox model.

Gouveia, João Pedro, Palma, Pedro, Simoes, Sofia G., 2019. Energy poverty vulnerability index: A multidimensional tool to identify hotspots for local action. Energy Rep. 5, 187–201.

Gouveia, João Pedro, Seixas, Júlia, Mestre, Ana, 2017. Daily electricity consumption profiles from smart meters - Proxies of behavior for space heating and cooling. Energy 141, 108–122.

Kontokosta, Constantine E., Tull, Christopher, 2017. A data-driven predictive model of city-scale energy use in buildings. Appl. Energy 197, 303–317.

Kwac, J., Flora, J., Rajagopal, R., 2014. Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid 5 (1), 420–430, Conference Name: IEEE Transactions on Smart Grid.

Langevin, J., Reyna, J.L., Ebrahimigharehbaghi, S., Sandberg, N., Fennell, P., Nägeli, C., Laverge, J., Delghust, M., Mata, É., Van Hove, M., Webster, J., Federico, F., Jakob, M., Camarasa, C., 2020. Developing a common approach for classifying building stock energy models. Renew. Sustain. Energy Rev. 133, 110276.

Oliveira Panão, Marta J.N., Brito, Miguel C., 2018. Modelling aggregate hourly electricity consumption based on bottom-up building stock. Energy Build. 170, 170–182.

Österbring, Magnus, Mata, Érika, Thuvander, Liane, Mangold, Mikael, Johnsson, Filip, Wallbaum, Holger, 2016. A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model. Energy Build. 120, 78–84.

Rasmussen, Christoffer, Bacher, Peder, Calì, Davide, Nielsen, Henrik Aalborg, Madsen, Henrik, 2020. Method for scalable and automatised thermal building performance documentation and screening. Energies 13 (15), 3866, Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.

Romero Rodríguez, Laura, Sánchez Ramos, José, Guerrero Delgado, MCarmen, Molina Félix, José Luis, Álvarez Domínguez, Servando, 2018. Mitigating energy poverty: Potential contributions of combining PV and building thermal mass storage in low-income households. Energy Convers. Manage. 173, 65–80.

Swan, Lukas G., Ugursal, V. Ismet, 2009. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renew. Sustain. Energy Rev. 13 (8), 1819–1835.

Voulis, Nina, Warnier, Martijn, Brazier, Frances M.T., 2018a. Understanding spatio-temporal electricity demand at different urban scales: A data-driven approach. Appl. Energy 230, 1157–1171.

Voulis, N., Warnier, M., Brazier, F.M.T., 2018b. 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. pp. 1–6.

Wang, Zhe, Hong, Tianzhen, Li, Han, Ann Piette, Mary, 2021. Predicting city-scale daily electricity consumption using data-driven models. Adv. Appl. Energy 100025.