IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 28, 2018, accepted October 31, 2018, date of publication November 15, 2018,
date of current version December 18, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2881413

EMPOWERING, a Smart Big Data Framework for Sustainable Electricity Suppliers

GERARD MOR[1], JORDI VILAPLANA[2], STOYAN DANOV[1], JORDI CIPRIANO[1], FRANCESC SOLSONA[2], AND DANIEL CHEMISANA[3]

[1]CIMNE-BEE Group Lleida, 25002 Lleida, Spain
[2]Department of Computer Science & INSPIRES, University of Lleida, 25001 Lleida, Spain
[3]Department of Environmental and Soil Sciences & INSPIRES, University of Lleida, 25198 Lleida, Spain

Corresponding author: Francesc Solsona (francesc@diei.udl.cat)

This work was supported in part by the Intelligent Energy Europe (IEE) Programme and the Ministerio de Economía y Competitividad under Contract TIN2017-84553-C2-2-R and in part by the European Union FEDER (CAPAP-H6 network) under Grant TIN2016-81840-REDT.

ABSTRACT This paper presents the EMPOWERING project, a Big Data environment aimed at helping domestic customers to save electricity by managing their consumption positively. This is achieved by improving the information received about energy bills and offering online tools. The main contributions of EMPOWERING are the creation of a novel workflow in the electricity utility sector regarding the implementation of data analytics for their customers and the fast implementation of data-mining techniques in massive datasets within a Big Data platform to achieve scalability. The results obtained show that EMPOWERING can be of use for customers of electrical suppliers by changing their energy habits to decrease consumption and so increase environmental sustainability.

INDEX TERMS Big-data, electricity supply industry, sustainable development, domestic consumption.

I. INTRODUCTION

The built environment sector is becoming the leading consumer of energy in the world, accounting for 40% of global energy use and one third of overall greenhouse gas emissions [1]. Within the built environment, in 2015, residential energy consumption amounted to around 25.4% of total final energy use in the European Union [2]. Therefore, to achieve the European 2020 targets, changes in the consumption patterns of EU households are urgent and necessary. To mitigate the energy and environmental pressures caused by household energy use, substantial research and development efforts have been made into energy-efficient technologies [3]. In recent years, improving energy efficiency and reducing energy demand have been widely regarded as the most promising, fastest, cheapest and safest ways to mitigate environmental pressures and climate change [4]. As a result, heating and cooling systems now use less energy than ever. However, final energy consumption has not decreased as expected. On the contrary, energy consumption has tended to increase. An analysis carried out within the EU-funded ODYSEE and MURE projects [5] quantified the increase in the energy efficiency of domestic appliances in Europe over the 2000 to 2012 period at 21% while the increase in final energy consumption was 75 Mtoe for the same period. One reason appears to be that much technology is made available to the public without adequate instruction and support. Although technological advances are significant for promoting energy conservation and improving energy efficiency [6], it is increasingly recognized that behavioral factors are of greater significance for energy conservation [7]. It has been suggested that behavioral changes can be just as effective as technological changes [8]. In [5], it was stated that changes in heating behavior had an impact on energy consumption by reducing it by 20 Mtoe over the over the period from 2000 to 2012. Since 2008, the level of this behavioural effect has doubled to 2.6 Mtoe/year, compared with 1.2 Mtoe before. Effective long-term strategies should engage people directly in efforts to reduce their energy consumption. This should be achieved through the implementation of environmental policies aiming at changing energy use behavior, as highlighted in [9]. Acknowledging people as an active element in the energy system should lead to efforts to better understand how people interact with energy and to stimulate the development of Energy Awareness services that attempt to change how and when people use energy.

Regarding the change of the energy behavior of consumers, in recent decades, many psychological models have been developed and adopted to explore how householders consume energy and the factors that influence this [10]. Different types of intervention strategies have been developed with the aim of stimulating changes in people's energy use behavior and thus achieving energy savings [11].

The overall aim of the EMPOWERING project is to empower consumers by involving, informing and helping them to take measures to save energy on the basis of the information they receive from their utility company. More specifically, the consumers' aim consists of achieving measurable energy savings.

The main contribution of EMPOWERING consists of a novel dataflow procedure for electric utility companies to standardize data communication, cleaning, storage and analysis. This workflow is based on secure API REST [12] communication, a set of ETL (Extract, Transform and Load) modules to clean and store the data in the EMPOWERING databases and a set of analytical modules to infer information from the energy consumption. EMPOWERING analyses data across the database of clients by making unsupervised learning searches and inferring clusters of similar types of domestic customers according to different information fields by means of data-mining techniques. This procedure can account for similarity between neighborhoods, size of building, number of occupants, climatic zone, etc. It provides a means to make comparisons of energy consumption with similar customers, namely between members of the same cluster. EMPOWERING offers specific, personalized, targeted information about whether one's consumption is above or below a cluster average over a season. This can show a need for space heating systems to be checked, or the building envelope to be improved. The large amount of data handled cannot be processed efficiently using traditional databases. These are the foundations of the smart Big Data framework developed within the EMPOWERING project.

The EMPOWERING services can deal with different data granularity, from monthly-based data coming from standard meters, to hourly-based data from smart meters. However, notable benefits are reached when hourly metering is used. For instance, alarms can be set up that detect abnormally high consumption levels for base-load appliances such as refrigerators or freezers. Some of these possibilities have already been developed within the EMPOWERING project with the collaboration of four electric utility companies in Europe, but the potential is far from the mainstream. The EMPOWERING project aims to accelerate the transition of the use of this type of service from pioneering companies to mainstream best practice.

II. RELATED WORK

Many data-mining techniques have been used to predict electricity consumption [13], [14]. These include neural networks (NN) [15], support vector machines (SVM) [16], support vector regression (SVR) [17], decision trees [18], auto regressive integrated moving average (ARIMA) models [19], clustering models [20], decomposition models, grey box models [21], and regression models [22]. Ahmad *et al.* [23] noted that NN and SVR have been used extensively for forecasting residential electricity consumption. Suganthi and Samuel [17] considered NN and SVR suitable for predicting industrial energy demand. They concluded that the two models have advantages and disadvantages and that it is inconclusive which is the best for energy forecasting. In [18], the performance of regression analysis models, decision trees, and NN for energy forecasting were compared. In the winter period, NN performed slightly better, whereas in the summer period, the decision tree model performed somewhat better than the other two. Peral *et al.* [24] presented a multidimensional hybrid architecture to make energy consumption predictions based on energy data-mining techniques that additionally makes use of current energy data enriched by external unstructured Big Data information. Predictive data-mining has been also applied to the building operation stage to predict its overall energy consumption [14]. Data-mining can be also used to obtain deeper insights into the data, to try to discover associations, correlations, and intrinsic data structures in Big-data. This is called descriptive data-mining. Compared with predictive data-mining, the descriptive version is more flexible in application, as it does not involve a training process and the knowledge of the discovery process is not guided by predefined targets. Descriptive data-mining has mainly been applied at the building operation stage for fault detection and diagnostics [25], [26]. Popular techniques include association rule mining, anomaly detection and clustering analysis.

Quilumba *et al.* [27] proposed a combination of predictive and descriptive data-mining procedures, recognizing the importance of differences in energy consumption patterns. They proposed a prediction approach based on clustering customers according to their consumption behavior and then predicting the energy consumption of the whole population by aggregating the forecasting of each single cluster. They applied this strategy to predict electricity consumption and demand for event-organising venues in the residential and commercial sectors. Clustering has also been used in the literature to group energy consumers with similar characteristics [28], [29] and to detect atypical, usually undesired, user behavior [30], [31].

The results of the studies [32], [33] show that a combination of statistical analysis with prediction models (holistic, simulation and inverse models), complemented in some cases with monitoring data analysis, can be a powerful tool for developing urban energy action aimed at reducing the energy consumption not only of existing buildings but also in higher geographical areas, such as neighborhoods or districts.

Big Data technology gives insights into how we think about a certain topic [34]. Big Data tools can manage structured, unstructured and semi-structured data [35]. Various data-acquisition Internet-of-Things (IoT) devices are penetrating into the wider world and are able to collect information spanning different areas [36]. The estimated installed base of smart meters worldwide will surpass 1.1 billion by 2022 [37], and will collect electricity usage data in the range of 15 minutes each. This is up to a three thousand-fold increase in the amount of data utilities processed in the past. It means that by 2022 the electric utility industry will be swamped by more than 2 petabytes of data annually from smart meters alone. Cisco [38] estimated that the data generated by devices would reach 507.5 zettabytes (ZB) per year (42.3 ZB per month) by 2019. This immense growth of data cannot be processed efficiently using relational databases.

III. EMPOWERING

A. ARCHITECTURE

Fig. 1 shows the general architecture of the EMPOWERING system, which is designed to tackle the following IT challenges: (i) to provide a means to link the local utility database to the Big Data analysis environment, (ii) to offer high quality in the delivered services, (iii) to provide batch-processing data analytics services and (iv) to ensure data privacy and security.

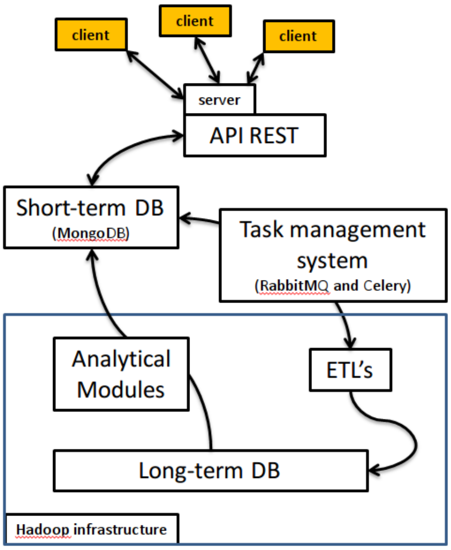

FIGURE 1. **EMPOWERING architecture.**

The Big Data architecture developed within EMPOWERING is a Representational State Transfer (REST) framework which provides an Engine with a technology aware interface.

A REST style architecture conventionally consist of a client-server paradigm. REST's client-server separation simplifies component implementation and allows intermediary modules, like proxies, gateways, caching systems and firewalls to be inserted into middle levels without changing the interface between the main components.

This architecture allows the storage and wrangling of large amounts of data. This is made up by a combination of low-cost hardware and database technologies that allows the acquisition, allocation and extraction of data to be processed in a distributed cluster. Essentially, the storage is split into **Short-term** and **Long-term** databases (DB), which have different characteristics according to the quantity, type and usage of data stored in them.

The EMPOWERING Big Data framework is entirely developed using open-source software. It is mainly composed of 3 components: *API REST*, *Task Management System* and the *Hadoop infrastructure*.

1) API REST

this is the communication interface between the server and the Client REST, and thus also with the utilities. The Application Program Interface (API) is fully developed following the REST standard. This component is the utilities' gateway to communicate and configure the Engine. The aim is to enable a Service-Oriented Architecture (SOA), offering specialized energy services to the customer and the utility system administrators. This is not a simple issue. The main objectives of the API are (i) to set and configure the services (see section III-C), (ii) to import data into the Engine and (iii) to export data from the Engine. These objectives are addressed using different technologies. Data import and export are enabled using the *Eve* framework to implement the Web service. *MongoDB* is the technology used for the short-term DB. It is buffer storage for data reception and sending in fast environments. It is the data storage directly connected to the API and provides high communication bandwidth. It supplies temporary storage, acting as a cache memory, prior to permanent storage in the long-term database. *ExtJS* technology was used to implement User Interface (UI) for setting and configuring the services. *OpenAM* provides open source Authentication, Authorization, Entitlement and Federation software. The *Flask* and *Python* modules implement all the server functionalities in order to deploy a web API server. Flask allows customizable, fully featured REST Web Services to be built and deployed effortlessly, which greatly simplifies the configuration of the API.

2) TASK MANAGEMENT SYSTEM

this level is in charge of scheduling and synchronizing the tasks in the engine by means of *RabbitMQ* and *Celery*. In essence, the scheduler picks up the new task to be executed in EMPOWERING according to a scheduling policy. The FIFO policy was chosen because the batch operation of the tasks made other variants (like Round Robin), frequently applied in time-sharing environments, inefficient. Celery is the scheduler itself. RabbitMQ is a fast internal message-queuing system used to interchange information between tasks with different paradigm technologies.

3) HADOOP INFRASTRUCTURE

Apache Hadoop is an open-source framework that provides tools for distributed storage and processing. It allows organizations to process and analyze large volumes of unstructured and semi-structured data, heretofore inaccessible, in a cost- and time-effective way.

Apache **Ambari** is used in order to manage the Hadoop cluster. It allows nodes to be added and removed, new components to be installed in existing working nodes, the cluster monitored, etc.

The two main Hadoop components are YARN and HDFS:

- HDFS (Hadoop Distributed File System) consists of slave components called DataNodes where data is physically saved and a master process called NameNode that is responsible for mantaining the file system directory tree and has the information of where data effectively is (i.e. which blocks are available in every DataNode). All HDFS reads and writes are managed by DataNode.
- YARN (Yet Another Resource Negotiator) is responsible for processing Map-Reduce tasks using the master-slave paradigm. It consists of the ResourceManager (master similar to NameNode). It is in charge of managing the launched tasks. The NodeManager resides in the slave nodes. It receives Map or Reduce orders from the ResourceManager and executes those tasks in YARN containers.

There are many high level applications running on the main components. Two of them were used in this project:

- *Hbase*: Distributed key-value database. Provides real-time read/write access and is built on top of HDFS. Hbase is used as the long-term big-data DB. It is formed by hundreds of thousands of AMI (Advanced Metering Infrastructure) devices used in EMPOWERING.
- *Hive*: Data warehouse on top of HDFS which provides SQL-like querying which are translated into MapReduce functions.

The YARN component is recursively used when Extract, Transforms and Load (ETL), and analytical modules are running. Initially, multiple asynchronous ETL functionalities aggregate, clean and transfer the data from the short-term to the long-term DB. These functions pre-process the input data to ensure the quality and format of the long-term DB.

Once the information is stored in the long-term DB, asynchronous analytical modules are implemented to generate the needed results for the services offered to the utility. The technologies used for the algorithms are a combination of *R*, *Hive*, and *Python software libraries* using the Map Reduce [39] paradigm to allow complex calculations over large sets of data. R is an open-source programming language for statistical computing. In order to use R in the Hadoop environment, the Rhipe and Rhadoop packages were used. These packages offer access to the long-term DB and facilitate the implementation of Map Reduce algorithms using common R functionalities. Python can also be used in the same manner with the MRjob, Happybase and Snakebite libraries. Python scientific libraries, such as Pandas, SciPy or NumPy, enable other advanced means for data analysis as an alternative to R. Hive is a data warehouse system for Hadoop. It provides functionality for data summarization, querying, and analysis of data. Hive queries are written in HiveQL, an SQL-like language.

The combination of these languages allows the use of the most highly optimized implementations according to the requirements of the algorithm and this generates less development effort and a shorter data processing time when the code is executed.

B. DATA

EMPOWERING services mainly rely on three categories of data: (1) energy consumption and contract, (2) end-user's and (3) third-party data.

1) ENERGY CONSUMPTION AND CONTRACT DATA

this is the kind of data used for billing (e.g. consumption data, contract details). This encompasses consumption data, either read at a low frequency manually, or by analogue meters, or estimated (quarterly, bi-annually, annually, etc.), as well as fine consumption data from smart meters (sub-hourly, hourly, daily, monthly, etc.). A certain type of consumption data may require clients' consent to collect or display. Thus, this type of data may not be available for all customers.

2) END-USER DATA

this type of data is not directly accessible by the customers because it does not serve for billing purposes. It is usually collected via online forms or surveys. Services relying on this type of data depend on the willingness of customers to fill in information about their dwellings and equipment. It can be erroneous or incoherent, so services based on this information have to consider data inaccuracies.

3) THIRD-PARTY DATA

these data is obtained from remote databases or provided by third parties and do not concern the user directly. This can be meteorological, statistical, etc.

EMPOWERING was conceived as a Big Data ICT architecture because of the large amount of data to be managed. More specifically, in the first services implemented from 2013 to 2016, the EMPOWERING architecture was managing 3 years of historical data from 70,000 contracts with the end users of two European electricity trading companies on an hourly basis and 30,000 contracts on a monthly basis, altogether corresponding to 1,831 million measurements of electricity consumption.

C. SERVICES

This section describes the EMPOWERING services. These constitute the main outputs of the analytic modules and are delivered to the final user in multiple formats and timescales (i.e. web, reports).

In addition to the usual services currently provided by electric utilities, such as consumption billing or historical monthly consumption, others seek to increase the benefits and volume of useful information that reaches the end users. These innovative services focus on the following topics: weather-normalized consumption comparisons compared with similar consumers, personalized energy-saving tips, tariff comparisons, consumption prediction and consumption alerts. These are the most relevant services currently developed in EMPOWERING. Most of them are based on one or multiple data-mining techniques to detect the weather-dependent share of consumption, clustering similar neighbors or forecasting the energy consumption.

1) WEATHER-DEPENDENCE ANALYSIS

this can be understood as a pre-treatment service. It is widely used in many services, e.g. normalized benchmarks, clustering of similar neighbors and consumption prediction or alerts. It estimates customers' energy consumption with respect to the weather at their locations. These services use several linear-regression techniques to correlate the energy consumption for space heating or cooling with the outdoor temperature. Households with strong weather dependence are associated with higher consumption levels in winter and higher outdoor temperatures in summer. Fig. 2 depicts the information provided to the customers: monthly consumption and the average monthly evolution of temperature over the preceding 12-month period. An explanatory text is attached so that the customers can understand the correlations between their energy consumption and the outdoor temperature better. In this case, it seems that electricity consumption is not weather dependent. Thus, this customer's consumption was similar throughout the year.

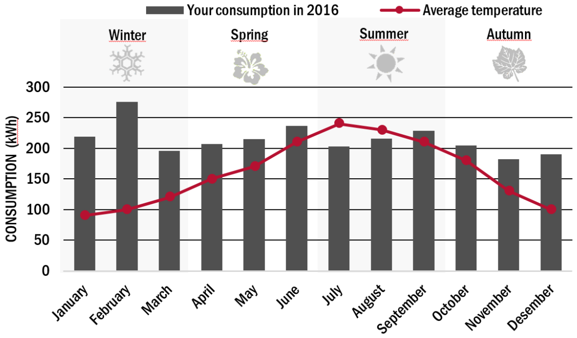

FIGURE 2. **Monthly consumption and average temperature over the last 12 months.**

The models used to determine weather dependence differ depending on the data granularity of the consumption data:

- **Monthly data**. For monthly data, a linear regression is used to fit the energy consumption with the degree of heating or cooling days in each month. The linear regression during heating and cooling periods expressed in equations 1 and 2 respectively is:

$$E_t = \alpha + H_h * (HD_t) + \epsilon, \quad (1)$$

$$E_t = \alpha + H_c * (CD_t) + \epsilon, \quad (2)$$

  where E_t is the electricity consumption at month t (Wh), α is the estimated baseload consumption (Wh), H_h is the estimated Heat Transfer Coefficient of the dwelling (not considering the performance of the systems) (Wh/K), H_c is the estimated Cool Transfer Coefficient of the dwelling during the cooling period (not considering the performance of the systems) (Wh/K), HD_t and CD_t are respectively the heating and cooling days in month t (K), and ϵ is the residual error (Wh), assumed to be white noise.

  The model parameters are estimated using the least-squares minimization approach. Customers with an estimated HTC higher than 100 Wh/K are assumed to have weather dependence during the cooling or heating periods. The adjusted coefficients α and HTC are also used to weather-normalize the monthly consumption when this information is used to compare historical consumption. In this case, the HD_t and CD_t considered correspond to the values of the last 12 months.
- **Hourly data.** Three-parameter (3P) and five-parameter (5P) models are used to estimate the relation between the daily aggregated electricity consumption and the average daily outdoor temperature. The 5P model is appropriate for modeling energy consumption data that include both heating and cooling, e.g. dwellings with a heat pump installed. 3P models are appropriate for modeling the electricity use in residences with a weather dependence during one of the periods (cooling or heating), e.g. dwellings with an electric chiller or boiler installed. The 5P model is presented in Equation 3.

$$E_t = \alpha + H_c * (T_t - T_c)^+ + H_h * (T_t - T_h)^- + \epsilon, \quad (3)$$

  where E_t is the electricity consumption at day t in Wh, T_t is the average daily temperature at day t (K), T_c is the cooling change point temperature (K), T_h is the heating change point temperature (K). α, T_c and T_h are stimulated.

For the time-dependent consumption comparison modules, a weather normalization analysis is applied. It consists of estimating the actual consumption by considering the ratios between the HD_t or CD_t from the previous and current periods. This estimation allows the comparison of the electricity consumption for different periods on a basis of similar weather. Once weather normalization has been applied, the differences in energy consumption between both periods are considered to be due to other factors (user behavior, new appliances, etc.).

2) CLUSTERING OF SIMILAR CUSTOMERS

energy consumption can be compared either against historical customer consumption or with other customers with similar characteristics. These consumption comparisons also take into account different time periods: daily, monthly, quarterly, semi-annual, yearly, bi-annual and triennial. In order to obtain similar customers, a clustering procedure is performed.

Several data-mining techniques are used, ranging from supervised learning approaches, based on similar contract information (contracted power, tariff) or geographical information (municipality, postal code, region), as K-nearest neighbours, to unsupervised learning algorithms such as Self-Organizing Maps (SOM) and K-means. The selection of the best grouping criteria for each customer is based on an optimization procedure which is aimed at minimizing a cost function (Equation 4) that consists of the difference between the monthly electricity consumption of a customer and the average of their peers and the dispersion of this monthly consumption within this group. To increase the robustness of the grouping criteria selection, the optimization procedure considers the last 12 months available.

$$\min_X f(X,c) = \frac{1}{n}\sum_{i=1}^{n} |E_{ic} - \overline{E_{iX}}| + \frac{Q_3(E_{iX})) - Q_1(E_{iX}))}{Q_3(E_{iX})) + Q_1(E_{iX}))}, \tag{4}$$

where X = similar customers, c = customer, n = number of months, $Q_3(E_{iX})$ and $Q_1(E_{iX})$ are the i*th* monthly consumption 75% and 25% percentiles of the similar users.

In general, the meaningfulness of the grouping criteria is related to the availability of input data and their characteristics. For instance, in the case of customers with only consumption data and no contract or survey information available, the meaning of the chosen grouping criteria is only related to the range of yearly consumption or the shape of the yearly profile. In other cases, when more contract information is available, the meaning of the best grouping criteria could be related to similar contracted power, heating or cooling dependencies, weather severity and also consumption indicators.

In the case of the unsupervised approach, a clustering algorithm is used to group the different kinds of customer. The first step is to train an SOM, a Neural Network (NN) that makes up the low-dimensional representation of the overall set of customers. When some customer features are clustered in a specific neuron, that represents a similar group of customers. The next step consists of a second clustering procedure, using the K-means technique, to find the emergent structures. The inputs used in this second clustering are the centroids from the SOM neurons and their mapping position. The emergent structure offers a more abstract description of a complex system consisting of low-level individuals. The number of groups for the K-means algorithm is optimized using the Gap Statistic index [40].

Two types of features are used in the training phase of the clustering algorithm. The first considers static customer features, such as contract information (contracted power, tariff, heating or cooling resources, location or yearly consumption), weather dependence indicators (explained in section III-C1) and daily or weekly consumption profiles averaged over a long period. The second one considers customer dynamic features. It consists of determining likely cyclical consumption patterns. Daylight imposes a natural rhythm on human behavior, making daily series an obvious choice. Inferring how consumers use electricity during different periods of the day on different days of the week and seasons of the year, is considered the most relevant information to be found. In order to obtain this, the SOM + K-means algorithm is used to detect patterns of daily consumption series over the whole set of customers. Fig. 4 depicts a subset of those detected patterns for a utility of 6,500 customers. This information is the used by a classification algorithm (i.e. SVM) that detects the pattern closest to every real daily consumption series of each customer. Thus, the results of this classification are a discrete time series of the closest consumption pattern over time. With this discrete time series, the signature of daily consumption series over a limited period (3-4 months, at least) is calculated. Finally, the signature and a K-means algorithm are used to detect the groups of similar customers in terms of energy behavior, because users with similar user behavior over time seem to have a similar signature for daily consumption over a time period.

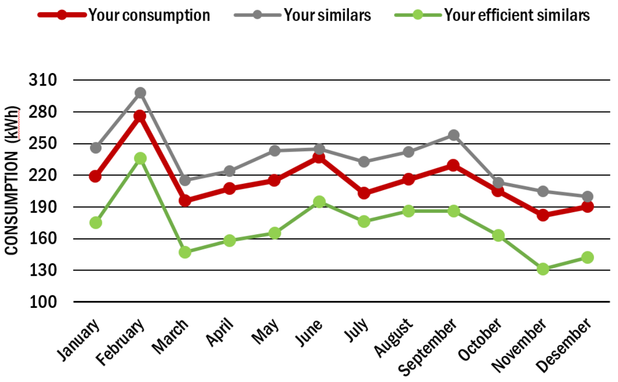

FIGURE 3. **Monthly consumption over the last 12 months compared with similar users.**

Fig. 3 shows a comparison of consumption over the previous 12 months between a customer, similar customers and the most efficient customers within the corresponding cluster.

3) FORECASTING

electricity forecasting is widely used in EMPOWERING to give consumption prediction information to customers and the utility. The techniques used for forecasting depend highly on the customer characteristics and their energy usage. The AutoRegresive Integrated Moving Average with eXhogeneous variables (ARIMAX) is used for those contracts that are weather dependent, because multiple independent variables could be considered in addition to the lagged consumption time series, e.g. the outdoor temperature, solar radiation or wind speed. In the case of contracts without weather dependence, Generalized Additive Models with Autoregressive fitting of the Residuals (GAMAR) are used. Alternatively, the consumption of this type of customers could be forecast using a decision tree which is trained by the information inferred in the clustering of similar customers in order to make the day-ahead predictions.

4) TARIFF COMPARISON

this set of services summarizes high-frequency energy consumption measures and integrates them into the tariff

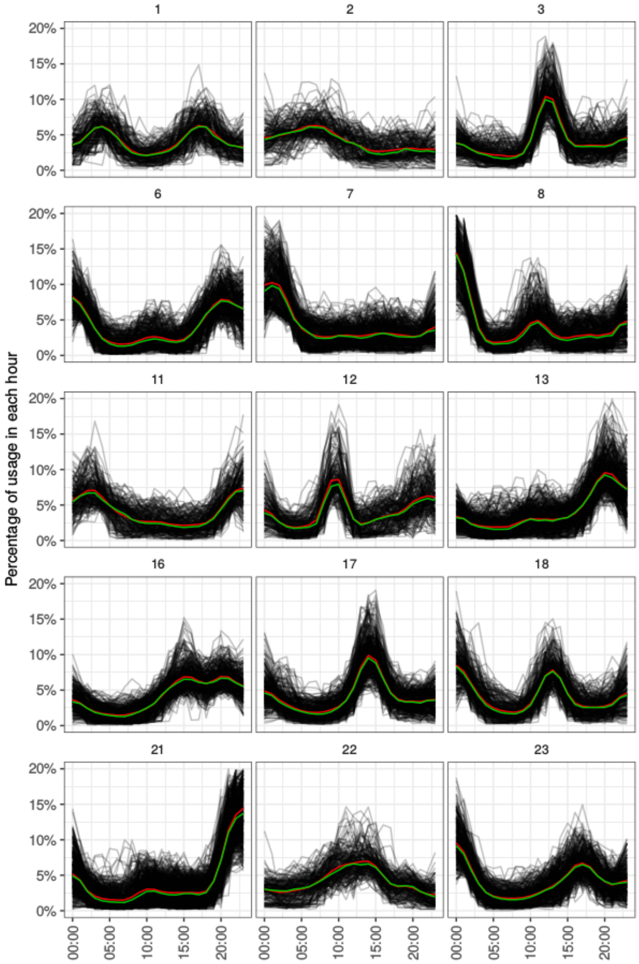

FIGURE 4. Subset of representative daily consumption series of all the customers.

information of the utility. Thus, the customer can visualize which tariff is the most cost-effective considering their real energy usage. Fig. 5 depicts the result of the daily consumption of a customer considering a time-of-use tariff in a single month.

5) PERSONALIZED ENERGY-SAVING TIPS

these are the most important services for energy awareness. The energy-saving tips are delivered once a month and the methodology used to select them differs depending on the data frequency. For the monthly data, the energy-saving tips are related to each customer's weather dependence (as defined in III-C1) and the season of the year. When a customer has strong weather dependence, they will receive tips related to space heating or cooling systems. In the case of customers with smart meters, the selection of tips is done following the procedure explained in Algorithm 1. This procedure is performed once a month. To avoid repetition, the selected tips are excluded from the procedure for a period of four months.

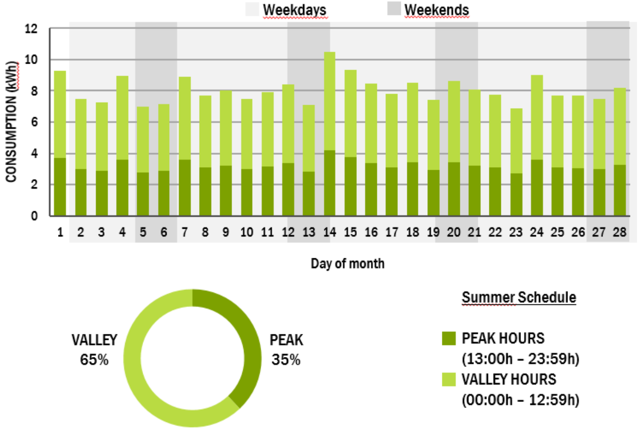

FIGURE 5. Daily consumption for each of the two tariff periods contracted by a customer in one month.

Algorithm 1 Energy Saving Tips

```
1: Clustering of users into similar groups based on each customer's daily consumption pattern and following the same techniques as in Section III-C2.
2: Evaluation of the average daily pattern of each cluster and the hourly percentage difference between this averaged pattern and each customer's pattern.
3: Definition of a set of around 100 energy-saving tips and weighting of each tip every hour of the day. For example, tips linked to cooking have a higher weight at midday and in the evening.
4: Calculation of the accumulative product between the hourly weight of each tip and the hourly percentage difference of each customer's pattern.
5: Classification of the tips to be delivered to each customer based on the score obtained.
6: Delivery of the three tips with the highest scores.
```

6) ALERTS AND ALARMS

in order to avoid over-consumption in upcoming bills, these services calculate the bias between the actual consumption of each customer and their historical consumption and set up an alarm. This allows customers to react within the period between two consecutive energy bills. The frequency of the alarms is directly dependent on the data granularity. For instance, daily or weekly consumption data is needed for monthly alarms. The platform delivers the alerts to the customers through visual interfaces and direct messages.

IV. RESULTS

A. EVALUATION METRICS

The evaluation of energy savings was based on the difference-in-difference multi-parameter linear regression method according to [41] and re-arranged in Equation (5). This method is widely adopted to evaluate the behavior of energy-efficiency based programs. It only evaluates the differences caused by the delivering of the EMPOWERING services, avoiding the rest of the external factors that affect the customers. It was implemented as a service within the analytical tool with access to the long-term databases. The EMPOWERING databases contain consumption data for all customers, classified as customers who receive the EMPOWERING services, EMPOWERING Group (*EG*), and the remaining ones, making up the Non- EMPOWERING Group (*NG*). An extension of the EMPOWERING data model, within the API Restful, was implemented to include each customer evaluated in the corresponding group and the date when they started using the EMPOWERING services. The energy savings analysis is calculated for each group according to the Averaged Daily Consumption (*ADC*). The ADC is obtained as the ratio of the aggregated monthly consumption. Once the *ADC* of each customer, month and period has been determined, the customer is inserted into a group. The relationship between these variables can be found as a multiple linear regression model and are used to find the *ADC* (Equation (5)).

$$ADC = \alpha + \beta * G_E + \gamma * t + \delta * (G_E * t) + \varepsilon, \qquad (5)$$

where:

α Independent parameter. It could be assumed to be the theoretical base-load average daily energy consumption of the total number of customers.

β Parameter related to the difference in energy consumption caused by the effect of belonging to the (*EG*) or (*NG*) groups.

G_E Treatment variable. $G_E = 1$ if the customer belongs to *EG* and $G_E = 0$ if the customer belongs to *NG*.

t Time period variable. $t = 1$ if the month falls within the evaluation period and $t = 0$ if it is outside the evaluation period.

γ Parameter related to the time trend effect.

δ Parameter related to the combined effect of the customer group and the time trend.

ε Uncertainty error accounting for all effects not considered in the model.

The parameters are determined through a least square minimisation of the residuals. Once all the parameters have been determined, the expected energy savings, (E_S), achieved by the customers belonging to *EG* is determined with Equation 6.

$$E_S = \frac{\delta}{(\alpha + \beta + \gamma)} * 100\% \qquad (6)$$

B. ENERGY SAVINGS

The EMPOWERING architecture and services were applied in three pilot experiments in France, Spain and Austria for slightly over 2 years, from November 2013 to December 2015. In each country, a local electricity-supplier was responsible for gathering data from customers, putting this into the analytical platform presented in section III-A to obtain data analysis and deliver them to the customers through several user interfaces. The details of the communication channel to deliver the services and the number of users included in the *EG* and *NG* groups as follows:

- **Spain.** The services were provided to the customers in two ways: (group 1) through an on-line portal and (group 2) as a monthly energy report. Meter readings were taken daily. The energy reports were sent together with the energy bill every 2 months. The *NG* and *EG* groups consisted of 3,129 and 1,582 customers respectively.
- **France.** The services were also offered to customers as an on-line tool within the utility web portal. Meters were read at a frequency of 6 months, but the services used an estimated 3-month consumption. To evaluate the energy savings of similar users, a clustering of the customers belonging to the *NG* was performed based on the contracted power: (group 1) low contracted power and electricity use limited to home appliances; (group 2) high contracted power and electricity used mainly in space heating systems; (group 3) low contracted power with occasional use of electricity for domestic hot water and space heating. The *NG* and *EG* were formed of 4,632 and 60 customers respectively.
- **Austria.** The services were offered to the customers as an on-line tool within the utility web portal. Meters were read every 15 minutes. The *NG* and *EG* groups were made up of 45,423 and 115 customers respectively.

After the test, the evaluation of the energy savings achieved by the group of customers belonging to *EG* was performed following the methodology defined in Eq. 6. Fig. 6 shows the percentage of energy savings achieved in the Spanish, French and the Austrian pilot projects. As can be seen, the customers who used electricity mainly for space heating and domestic hot water systems or those who received both energy reports and access to on-line tools achieved greater savings in electricity.

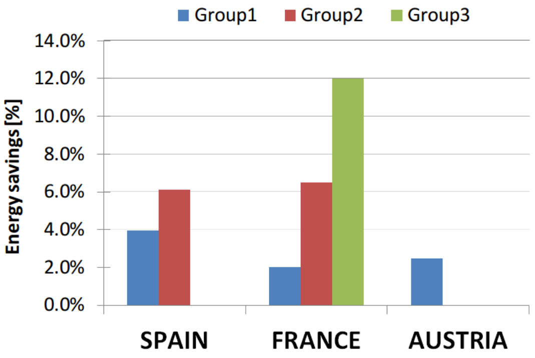

FIGURE 6. Average energy savings achieved in the Spanish, French and Austrian pilot projects.

Figure 7 shows the energy savings of the two groups of customers in the Spanish pilot project segmented into percentiles of electricity consumption. A similar pattern can be appreciated for both groups. In general, higher savings were achieved in the higher energy consumption segments. Both groups reduced consumption significantly. The saving was considerably higher for the customers using both the billing and on-line tools (11%), compared to the users who only

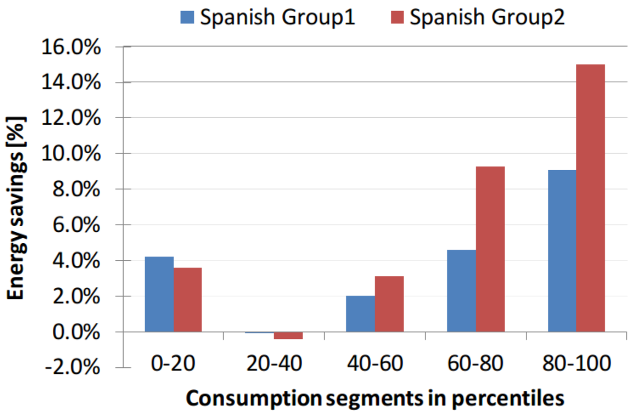

FIGURE 7. Energy savings per consumption distribution percentile range for the Spanish pilot project.

used the billing tool (6%). Thus, offering the services through multiple channels and the customers' interactivity with the portal improve the savings significantly.

Fig. 8 shows the energy savings achieved by the three groups of customers in the French pilot project, segmented into the percentiles of electricity consumption. It indicates higher energy savings for the middle-high percentile range customers in groups 1 and 2. In group 3, the savings are present over the whole range of consumption, with higher (up to 50%) savings for the largest consumers. It can also be seen that electricity savings were achieved in all three groups. Savings were higher in the groups where the electricity was used for both space heating and hot water (Groups 2 and 3). Considerable savings, above 20%, were achieved in Group 3, this being the group with more opportunities to modify their energy usage habits. The number of customers in the *EG* was relatively small for the three evaluation groups, allowing room for large uncertainty in the evaluated results.

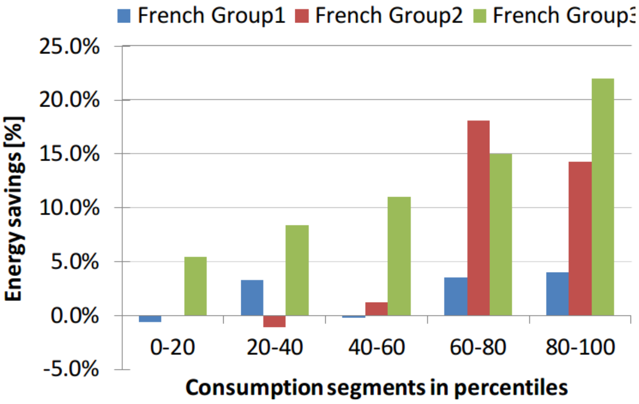

FIGURE 8. Energy savings per consumption distribution percentile range for the French pilot project.

Fig. 9 shows the energy savings achieved among the customers in the Austrian pilot project, segmented into the percentiles of electricity consumption. Higher energy savings were achieved by customers in the upper consumption segments, while the customers with lower consumption barely increased their energy consumption. The opt-in strategy also meant that only very motivated customers entered the portal. The baseline consumption before the services was 8.15 KWh/day and the average savings per user were 0.5 KWh/day.

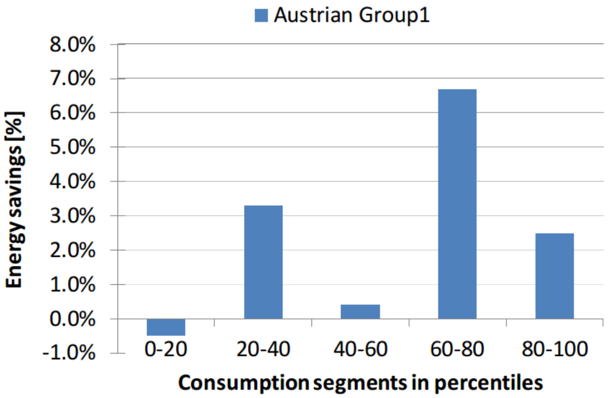

FIGURE 9. Energy savings per consumption distribution percentile range for the Austrian pilot project.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we present an efficient and scalable platform aimed at helping domestic customers to save energy by managing their energy consumption positively. The electricity savings achieved by using EMPOWERING ranged from 3 to 50% among the different pilot and user groups. Improvements in the behavioral aspects in energy use have considerable potential. The users' own motivation also seems to play an important role and thus, better results were achieved with customer involvement. The personal motivation for energy savings is based on different reasons and money saving is only one of them. Environmental concerns, governmental laws, social policies and technological restrictions are other powerful reasons where the future services should diversify in order to have greater impact. In addition, more encouraging and ad-hoc services must be provided to the final customers. Future work will analyze energy awareness depending on the nationality of the customers. We leave this for the future work due to the complexity of the diversity of features as well as clustering groups to be analyzed.

REFERENCES

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy Buildings*, vol. 40, no. 3, pp. 394–398, 2008.

[2] *Energy Balance Sheets—2013 Data*, Eur. Union, Brussels, Belgium, 2015.

[3] K. Zhou and S. Yang, "Understanding household energy consumption behavior: The contribution of energy big data analytics," *Renew. Sustain. Energy Rev.*, vol. 56, pp. 810–819, Apr. 2016.

[4] S. Sorrell, "Reducing energy demand: A review of issues, challenges and approaches," *Renew. Sustain. Energy Rev.*, vol. 47, pp. 74–82, Jul. 2015.

[5] L. Gynther, B. Lapillonne, and K. Pollier. (Jun. 2015). *Energy Efficiency Trends and Policies in the Household and Tertiary Sectors: An Analysis Based on the ODYSSEE and MURE Databases*. [Online]. Available: http://www.odyssee-mure.eu/publications/br/energy-efficiency-trends-policies-buildings.pdf

[6] L. Steg, L. Dreijerink, and W. Abrahamse, "Factors influencing the acceptability of energy policies: A test of VBN theory," *J. Environ. Psychol.*, vol. 25, no. 4, pp. 25–415, 2005.

[7] B. K. Sovacool, "What are we doing here? Analyzing fifteen years of energy scholarship and proposing a social science research agenda," *Energy Res. Social Sci.*, vol. 1, pp. 1–29, Mar. 2014.

[8] W. Prindle and S. Finlinson, "How organizations can drive behavior-based energy efficiency," in *Energy, Sustainability and the Environment*. Oxford, U.K.: Butterworth, 2011.

[9] G. T. Gardner and P. C. Stern, *Environmental Problems and Human Behavior*. Boston, MA, USA: Allyn & Bacon, 1996.

[10] T. Jackson, "Motivating sustainable consumption: A review of evidence on consumer behaviour and behavioural change," Centre Environ. Strategy, Univ. Surrey, Guildford, U.K., Tech. Rep., 2005.

[11] L. Steg, "Promoting household energy conservation," *Energy Policy*, vol. 36, no. 12, pp. 4449–4453, 2008.

[12] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, Dept. Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2000.

[13] K. Grolingera, A. L'Heureuxa, M. Capretza, and L. Seewald, "Energy forecasting for event venues: Big data and prediction accuracy," *Energy Buildings*, vol. 112, no. 15, pp. 222–233, 2016.

[14] B. Dong, C. Cao, and S. E. Lee, "Applying support vector machines to predict building energy consumption in tropical region," *Energy Buildings*, vol. 37, no. 5, pp. 545–553, 2005.

[15] F. J. Ardakani and M. M. Ardehali, "Long-term electrical energy consumption forecasting for developing and developed economies based on different optimized models and historical data types," *Energy*, vol. 65, pp. 452–461, Feb. 2014.

[16] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Appl. Energy*, vol. 123, pp. 168–178, Jun. 2014.

[17] L. Suganthi and A. A. Samuel, "Energy models for demand forecasting—A review," *Renew. Sustain. Energy Rev.*, vol. 16, no. 2, pp. 1223–1240, 2012.

[18] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.

[19] C. Fan, F. Xiao, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Appl. Energy*, vol. 127, pp. 1–10, Aug. 2014.

[20] W. El-Baz and P. Tzscheutschler, "Short-term smart learning electrical load prediction algorithm for home energy management systems," *Appl. Energy*, vol. 147, pp. 10–19, Jun. 2015.

[21] T. Berthou, P. Stabat, R. Salvazet, and D. Marchio, "Development and validation of a gray box model to predict thermal behavior of occupied office buildings," *Energy Buildings*, vol. 74, pp. 91–100, May 2014.

[22] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables," *Appl. Energy*, vol. 140, pp. 385–394, Feb. 2015.

[23] A. S. Ahmad *et al.*, "A review on applications of ANN and SVM for building electrical energy consumption forecasting," *Renew. Sustain. Energy Rev.*, vol. 33, pp. 102–109, Jul. 2014.

[24] J. Peral, A. Ferrández, R. Tardío, A. Maté, and E. de Gregorio, "Energy consumption prediction by using an integrated multidimensional modeling approach and data mining techniques with Big Data," in *Advances in Computer Science* (Lecture Notes in Computer Science), vol. 8823. 2014, pp. 45–55.

[25] C. Fan, F. Xiao, and C. Yan, "A framework for knowledge discovery in massive building automation data and its application in building diagnostics," *Autom. Construct.*, vol. 50, pp. 81–90, Feb. 2014.

[26] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Syst. Appl.*, vol. 42, pp. 4324–4338, Jun. 2015.

[27] F. L. Quilumba, W. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015.

[28] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, 2012.

[29] D. Takaishi, H. Nishiyama, N. Kato, and R. Miura, "Toward energy efficient big data gathering in densely distributed sensor networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 388–397, Sep. 2014.

[30] Z. Yang and B. Becerik-Gerber, "Modeling personalized occupancy profiles for representing long term patterns by using ambient context," *Building Environ.*, vol. 78, pp. 23–35, Aug. 2014.

[31] X. Li, C. P. Bowers, and T. Schnier, "Classification of energy consumption in buildings with outlier detection," *IEEE Trans. Ind. Electron.*, vol. 57, no. 11, pp. 3639–3644, Nov. 2010.

[32] R. Hobday, "Energy-related environmental impact of buildings," Tech. Synth. Rep. Annex 31, Int. Energy Agency Energy Conservation Buildings Community Syst. Programme, 2010.

[33] Z. Yu, B. C. M. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, "A systematic procedure to study the influence of occupant behavior on building energy consumption," *Energy Buildings*, vol. 43, no. 6, pp. 1409–1417, 2011.

[34] Big Data Research @ EVRY. *Big Data in Banking for Marketers. How to Derive Value From Big Data*. Accessed: Jun. 5, 2017. [Online]. Available: https://www.evry.com/globalassets/insight/bank2020/bank-2020—big-data—whitepaper.pdf

[35] W. Liu and E. K. Park, "Big data as an e-health service," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, 2014, pp. 982–988.

[36] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Syst. J.*, vol. 10, no. 3, pp. 888–900, Sep. 2016.

[37] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, "Big data analytics in power distribution systems," in *Proc. IEEE Innov. Smart Grid Technol. Conf.*, Feb. 2015, pp. 1–5.

[38] P. Isley, "Cisco visual networking index: Forecast and methodology," Cisco Syst., San Jose, CA, USA, White Paper ID1458683795628678, 2015.

[39] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[40] R. Tibshirani, G. Walther, and H. Trevor, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc., B (Stat. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.

[41] B. D. Meyer, "Natural and quasi-experiments in economics," *J. Bus. Econ. Stat.*, vol. 13, no. 2, pp. 151–161, 1995.

GERARD MOR received the B.Sc. and M.Sc. degrees in building engineering and decision making from Universidad Rey Juan Carlos. He is specialized in building energy consumption simulations and data mining, developing algorithms and managing big data information from buildings and dwellings. Nowadays, he is in-charge of the development and implementation of energy user awareness services for the BEE Data platform.

JORDI VILAPLANA received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Lleida, Spain, in 2009, 2011, and 2015, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Lleida. His research interests include eHealth, mHealth, cloud, parallel and distributed processing, big data, and data analysis.

STOYAN DANOV received the Ph.D. degree in thermal engineering from the Polytechnic University of Catalonia. He is currently a Senior Researcher at the BEE Group with background in the field of mechanical and thermal engineering, renewable energies, and optimization of the energy use in industrial processes and buildings. He is interested in sustainable development, with extensive experience in the field of clean technologies: performing numerical simulations, development of new products and services employing ICT. He has participated in numerous international research and development projects, consultancy and technology transfer projects for private companies.

JORDI CIPRIANO received the M.Sc. degree in industrial engineering and engineering from the Polytechnic University of Catalonia and the Ph.D. degree in ICT from the University of Lleida. He has been the Director of the BEE Group since its foundation in 2001. His background comes from the application of numerical methods for the analysis of air flows and thermal heat transfer in active building components and urban environment. For the last years, he has centered in modeling the energy performance of buildings, BIPV components and in analyzing the user behavior of tenants in residential buildings. He has accumulated more than 12 years of experience in collaborating with municipalities and building owners through national and international cooperative research projects. He has participated in more than 25 EU funded projects. He is one of the founders and a member of the board of directors of the energy service company INERGY (www.inergybcn.com), promoted by CIMNE and RSM Gassó. He also acts as a representative of the CIMNE-BEE Group in the E2B and INIVE research network. He is also one of the Spanish representatives within the AIVC and he is participating in the Annex 58 of the IEA.

FRANCESC SOLSONA received the B.S., M.S., and Ph.D. degrees in computer science from the Universitat Autònoma de Barcelona, Spain, in 1991, 1994, and 2002, respectively. He is currently a Full Professor with the Department of Computer Science, University of Lleida, Spain. He is the co-founder of the Hesoft Group company. His research interests include distributed processing, cluster computing, co-scheduling, administration and monitoring tools for distributed systems, cloud computing, linear programming, big data, data analysis, social networks, and bioinformatics.

DANIEL CHEMISANA received the Ph.D. degree in physics from the University of Lleida, Spain, in 2009. He is currently a Full Professor with the Department of Environment and Soil Sciences, University of Lleida. He is the Coordinator of the research group Applied Solar Energy Section. His research interests include photovoltaics, electrical characterization, solar energy materials, CPV, and solar concentration.