Greenhome: a household energy consumption and CO2 footprint metering environment

Genoveva Vargas-Solar, Maysaa Khalil, Javier Alfonso Espinosa-Oviedo, José-Luis Zechinelli-Martini

GREENHOME: a Household Energy Consumption and CO_2 Footprint Metering Environment

GENOVEVA VARGAS-SOLAR, French National Centre for Scientific Research, LIRIS laboratory, France
MAYSAA KHALIL, University of Technology of Troyes, France
JAVIER A. ESPINOSA-OVIEDO, University of Lyon, ERIC laboratory, France
JOSÉ-LUIS ZECHINELLI-MARTINI, Universidad de las Américas Puebla, Mexico

This paper presents the GREENHOME environment, a toolkit providing several data analytical tools for metering household energy consumption and CO_2 footprint under different perspectives. GREENHOME enables a multi-perspective analysis of household energy consumption and CO_2 footprint using and combining several variables through various statistics and data mining algorithms. This elastic and multi-perspective analytics facility is an element of the originality of GREENHOME.[reviewer] To test GREENHOME, the paper reports on experiments conducted for modelling and forecasting energy consumption and CO_2 footprint in the context of the Triple-A European project.

Additional Key Words and Phrases: Smart grid datasets, energy consumption, CO_2 footprint, Big Data, Internet of Things

ACM Reference Format:
Genoveva Vargas-Solar, Maysaa Khalil, Javier A. Espinosa-Oviedo, and José-Luis Zechinelli-Martini. 2021. GREENHOME: a Household Energy Consumption and CO_2 Footprint Metering Environment. 1, 1 (December 2021), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Global warming, and its impending follow-ups, have become a major global issue. Scientists and governments have agreed that cleaner and sustainable solutions can help in reducing the impact of this phenomenon. Studies like [16, 33, 43, 58] have agreed that the energy sector, specifically the electric sector, impacts carbon dioxide (CO_2) emissions and can be regulated by public policies to reduce greenhouse gases emissions. According to EU statistics [8], buildings represent 40% of all energy consumption and 36% of CO_2 emissions in Europe due to the age of buildings in European cities. [48] shows that if the current energy consumption pattern persists, the world's energy consumption will increase more than 50% before 2030.

The concept of "smart building" has been introduced to address problems implied by this observation. The principle is to integrate *datification*[1] into buildings to optimise their usage in terms of comfort and energy. A smart building uses sensors and software for automating some processes like control lighting [5], climate [13], entertainment systems, and

[1]Technological trend turning many aspects of our life into data for comprehension and value extraction (https://en.wikipedia.org/wiki/Datafication).

Authors' addresses: Genoveva Vargas-Solar, genoveva.vargas-solar@liris.cnrs.fr, French National Centre for Scientific Research, LIRIS laboratory, France; Maysaa Khalil, mayssa.khalil@gmail.com, University of Technology of Troyes, France; Javier A. Espinosa-Oviedo, javier.espinosa-oviedo@univ-lyon2.fr, University of Lyon, ERIC laboratory, France ; José-Luis Zechinelli-Martini, joseluis.zechinelli@udlap.mx, Universidad de las Américas Puebla, Mexico.

appliances [30]. It may include home security [30] such as access control and alarm systems and occupancy measures [1]. Besides, the integration of smart measuring devices in a household via the Internet of things (IoT)[2] allows collecting information used for generating beneficial insights to increase energy efficiency in households and turn them into smart ones [64]. For example, the energy consumption environment is an advanced metering infrastructure (AMI) that measures, collects, analyses consumption, and communicates with metering devices according to a schedule or on request [18, 36].

Advances in sensing and data analysis technology open the possibility of using collected data to be processed and analysed to enhance the efficiency of the energy consumption, and energy grid [53]. Discovering, analysing and predicting energy consumption in buildings and households is an emerging research area. Academic and industrial works have proposed methods to predict energy consumption. Some do not consider smart meter data, like the work reported by [32], that proposes a two-stage long-term retail load forecast* model considering the residential customer's attrition. Others like [34], do use smart meter* data to forecast micro-grid settings to learn spatial information shared among interconnected customers, and to address the over-fitting challenges [19] to predict buildings energy consumption using time series data [42]. Studies and manifestos like [58] agree that it is essential to have a thorough understanding of sustainable energy consumption.[reviewer]

This paper presents the GREENHOME environment that provides tools to scientists for combining different variables to produce energy consumption models. These models give different energy consumption perspectives that can help to understand and compare them. The environment enables a multi-perspective analysis of household energy consumption and CO_2 footprint using and combining several variables through different statistics and data mining algorithms. To test GREENHOME, the paper reports on experiments conducted for modelling and forecasting energy consumption and CO_2 footprint in a household in Picardie, a region in the north of France, using energy data collected during one year[3] in the context of the Triple-A European project.

GREENHOME implements two machine learning methods to forecast energy consumption: the auto-regressive integrated moving average model (ARIMA) and the autoregressive model with exogenous terms (ARX). For the case study used in our experiments concerning the estimation of energy consumption of a house in Picardie, we used data collected hourly during 2018. Experimental results show that the ARX model is assessed with a better residential mean square error (RMSE) than the ARIMA model. We show that the performance of the model increases by adding exogenous variables. Both models report better performance than the naive forecast model-persistence method. Through experiments, we compared how they adapt best to the analysis of the datasets. The application of these models on other datasets may lead to different results. Beyond the pertinence of using one model or another, we show the importance of GREENHOME as an environment that facilitates comparing models on the same dataset with different analytical criteria but, in particular, the comparison of models and results across experiments of the same type, namely energy consumption analysis.

The remainder of the paper is organised as follows. Section 2 discusses the background of smart metering analytics (modelling and forecasting techniques). Section 3 introduces GREENHOME, our household energy consumption and CO_2 footprint metering and predicting system. Section 4 reports results on experiments regarding the data preparation phase of experiments run on data sets about household energy consumption. Section 5 reports the setting and results of a data science experiment regarding the computation of energy consumption and CO_2 footprint. Section 6 introduces our experiment regarding prediction models applied for forecasting energy consumption in a household. It describes

[2]See the definition in the glossary in appendix A. In the remainder of the section, terms tagged with a '*' are defined in the appendix.
[3]Dataset available at https://github.com/javieraespinosa/Triple-A-household-energy-dataset

results and discusses results. Section 7 introduces related work describing and comparing existing smart metering systems. Section 8 concludes the paper and discusses future work.

2 BACKGROUND

This section introduces the most relevant aspects to consider when dealing with big data processing for the smart grid. The study discusses the conditions in which data is collected in smart metering environments*. Then, aspects to consider for analysing data for answering questions about power load analysis* and forecasting, anomaly detection*, load profiling* and the associated architectures. This paper addresses the analysis and forecasting of energy consumption. This section studies aspects to consider for addressing the problem.

Exploiting big data in a smart grid is done by collecting data by smart metering tools and analysing data for profiling and forecasting power load. This collection and analysis are done through a reference architecture implemented by big data smart grid environments.

2.1 Big Data Processing for the Smart Grid

Big data proposes strategies to analyse, extract information, and deal with datasets that are too large or complex for traditional data-processing systems [31, 49]. There is a consensus about its characteristics described by the well-known V's model [31]: *volume*, *velocity*, *variety*, *veracity* and *value*. Other V's have been also considered like *validity* [31] to refer to the period during which data are representative, *valid* for a given use, and *visibility* [31] determining the point of view from which data are collected and processed.

In the smart grid context, analysed and extracted big data can be collected using different smart meters* installed in buildings to gather information regarding the energy and gas consumption, meteorological measures and residents' behaviour. The ability to extract valuable insights using big data processing can improve the efficiency of the smart grid, decrease consumption and maintain a production-consumption real-time balance.

Smart Grid Data Analytics. The survey reported in [47] queried people on top analytics initiatives*. It shows that system modelling, asset optimisation and outage management are the drivers in utility operational expenditures. The conditions in which the utility industry operates and its asset-intensive nature explains that the system modelling is on the top of the list. Smart meters* data contribute to the implementation of load management and forecasting in two aspects:

- Customer characterisation: The electricity consumption profile is related to the customer's socio-demographic status. This allows the classification of customers. Therefore, the point is to recognise socio-demographic information about customers from load profiles and predict the loads according to their socio-demographic classification. Different techniques, including fast Fourier transformation, sparse coding, and clustering, were used to classify customers. In addition, data like location, floor area, age of consumers, and several appliances may help in the classification.
- Demand response implementation (DR)*: DR has played a vital role in balancing the supply and demand for electrical load [14]. Bill rebates, redeemable vouchers, discounts are some incentive payments derived from DR programs. DR programs may lead to success only if these two factors are achieved: (i) how to operate DR resources which are mainly related to customers, energy market, devices and utility company; and (ii) how to measure DR performance. However, traditional baseline estimation cannot characterise uncertainties due to

their deterministic modelling. This deficiency often results in erroneous system operations and miscalculated payments that discourage participating customers [46].

Smart Grid Big Data Analytics Architectures. The main objective of big data analytics* is to explore and process data and transform it into meaningful information such as patterns of operation, alarm trends, fault detection, and control commands [47]. It uses techniques proposed in different domains like data mining, statistical analysis, machine learning and artificial intelligence (AI). Smart Grid Analytics uses data science processes for combining different solutions. Therefore it uses different technologies for managing, integrating and processing datasets, including Data warehouses DWH, large scale data processing frameworks (e.g., Hadoop) and real-time processing (stream computing) [29]. For instance:

- Data warehouses (DWH) are used for storage.
- Apache Hadoop is an open-source software library, a framework that allows for the parallel processing of large data sets across clusters of commodity hardware using simple programming models.
- Stream computing tools monitor millions of events in a specific time window to react proactively, and they are behaviour-based architectures where events are analysed in real-time and action performed and then stored in databases for further analytics.

The smart metering components can be deployed in the cloud [56] using multiple back-end services that communicate with the outside using three interfaces:

- Cloud gateway communicates with the sensors. It ingests device telemetry and ensures that the target devices reliably receive control messages.
- Web Application Server is responsible for house residents and administrators' interface. It provides a user interface necessary for data visualisation and device management, and monitoring. It is also responsible for securing these interfaces.
- Protocol Bridge provides the connection between the platform and an external platform. It translates between the standard application protocol and the protocol used by the external system.

Network-centric architectures, for instance, Service-Oriented Architectures (SOA) and resource aggregation and virtualization, are possible solutions to achieve flexibility and scalability in the grid control and monitoring infrastructure [17]. For example, Fenix[4] develops the concept of Virtual Power Plant to abstract and model the presence of a vast number of distributed energy resources.[reviewer].

Atat et al. in [3] propose a survey with a broad overview of data collection, storage, access, processing and analysis. The way big data converge with smart energy systems is that of architectures (e.g., cloud, fog, edge) providing storage and computing resources, and that of algorithms for processing and analysing data for modelling and predicting energy consumption in many different perspectives to uncover hidden patterns, unknown correlations and other helpful information.[reviewer]

When combined with artificial intelligence, machine learning, smart grid Big Data analytics architectures, will bring about new applications, services, and opportunities [3]. This strategy will help revolutionise the "smart planet" concept, where smarter water management, health care, transportation, energy, and food will radically transform people's lives.[reviewer]

[4] http://www.fenix-project.org

Energy consumption of big data environments.[reviewer]. Information and communications technologies (ICTs) can enable powerful social, economic and environmental benefits. However, ICT systems give a non-negligible contribution to world electricity consumption and CO_2 footprint. [reviewer]

The increasing number of devices and IoT systems enabling cyber-physical industrial IoT environments may consume substantial energy. Thus, the relevant energy efficiency issues have led to the proposal of energy-efficient architectures [51, 59] consisting of sense entities, RESTful services hosted networks, cloud servers, and user applications. [reviewer]

Lorincz et al. [33] analyse the costs for the global annual energy consumption of telecommunication networks, estimate the ICT sector CO_2 footprint contribution, and predicts energy consumption of all connected user-related devices and equipment between 2011–2030.

To reduce energy consumption at all levels of the stack, green wireless communications reason about the use of environmentally sustainable materials, occupying less land space, accompanying less electromagnetic pollution, together with waste recycling and reducing wastes, and cost reductions [57].[reviewer]

2.2 Forecasting Energy Consumption

Analysing time-oriented data and forecasting values using time series are classic problems that analysts face in the field of energy consumption [21]. The focus is on short to medium-term forecasts where statistical methods are helpful. Short-term predictions provide forecasting over days, weeks, or months to the future. Short-term forecast purpose is identifying, modelling and interpolating patterns and insights launched by historical data. The motive for forecasting in the electric consumption time series is that predictions are critical for various decision-making tasks like estimating carbon footprint, reducing energy consumption, etc. The forecast here is a quantitative forecast, where the model uses historical data and formally summarises patterns in data and statistically outcome a relationship between the previous records and the estimated ones.

Forecasting Models. There are mainly three groups of forecasting models: engineering, statistical and artificial intelligence models. A review of prediction methods can be found in [64] and [62]. Engineering methods are comprehensive methods that use the structural characteristics of the building in the form of physical principles and thermodynamic equations as well as environmental information like climate conditions and occupants' activities. However, these methods need fine-grained details about the structure and the thermal characteristics of the building that unfortunately are not always available [55].

Statistical methods use historical data to correlate between instance consumption and previous consumption and most influencing variables. Consequently, the quality and quantity of historical data possess a crucial role in developing the model. Regression models, conditional demand analysis (CDA), auto-regressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and Gaussian mixture models (GMM) are some examples of statistical models [12, 25, 27, 54]. The objective is to achieve energy efficiency and help stakeholders make decisions about different levels (region, city, quarter). The models are applied within data analytics*, and data science pipelines that can generate continuous insight out of data produced by sensing buildings and households [38]. It is believed that the data science approach can bring a new perspective to the study of energy efficiency in buildings and electric savings [64]. Thus, data science pipelines have been specialised in smart grid and smart metering analytics processes.

Forecasting Pipeline. The forecast process transforms a set of inputs into a set of outputs based on specific criteria. The outputs' set can be apprehended as a single result related to energy consumption per hour. The steps followed in the forecast process are: (i) data preparation* that includes problem definition, data collection*, anomaly detection*

and attribute engineering; (ii) data analysis that includes selecting and fitting the model; (iii) validating the model; (iv) deploying a forecasting model and finally (v) monitoring the forecast model performance.

Model Fitting, Selection and Validation. For a given prediction problem, choosing one or more forecast models is necessary and fitting the model to the data. Fitting is the process of estimating the model's parameters using different methods, especially the method of least squares [7]. It is essential to define the meaning of performance carefully. It is tempting to evaluate performance based on the fit of the forecasting on the historical data. Many statistical measures describe how well a model fits a given data sample.

When more than one forecasting model seems reasonable for a particular application, forecast accuracy measures can also be used to discriminate between competing models like using the one-step-ahead forecast errors:

$$e_t\ (l) = y_t\ -\ \hat{y}_t(t\text{-}1)$$

Where $\hat{y}_t$(t-1) is the forecast y_t made one in a prior period. Suppose there are n observations for which forecast has been made. Forecast accuracy standard measures are, for example, the mean error (ME), the mean square error (MSE) and the residual mean square error, defined as follows:

$$ME = \frac{1}{n} \sum_{t=1}^{n} e_t(l)^2$$

The mean square error:

$$MSE = \frac{1}{n} \sum_{t=1}^{n} |e_t(l))|^2$$

The residual mean square error: RMSE= $\sqrt{MSE}$

MSE and RMSE are estimates of the expected value of forecast error. Their values should be close to zero, meaning that the forecast technique produces unbiased forecasts. If the mean square error drifts away from zero, this can show that the underlying time series has changed in some fashion and that the forecasting technique has not tracked this change. Both MSE and RMSE measure the variability in forecast error. The variability should be small. RMSE is a direct estimator of the variance of the one-step-ahead forecast errors.

Selecting a model that provides the best fit to historical data generally does not necessarily result in the best forecast model. Focusing on the model that produces the best historical fit often results in over-fitting. In general, the best approach is to select the model that results in the smallest RMSE or MSE value when the model is run on top of data not used for the fitting process. This is done after splitting data, one for model fitting and the other for performing testing. It is called a cross-validation method.

Model validation is the process of evaluating the model chosen to determine how it is likely to perform in the desired application. The principle of the validation pipeline is getting new inputs for the model, different from the data used for testing and training. Therefore, the data used to build the final model usually come from 3 datasets: (i) training dataset that the model is initially fit on using a supervised method, (ii) the fitted model is used to predict the observation of the testing dataset where the estimation error is calculated for evaluating the model and (iii) the validation dataset used to provide an unbiased evaluation of a final model fit on the training dataset. This final dataset can stem from the initial dataset or another one. A dataset that has never been used for training a model is called the holdout dataset. Data splitting is used here; generally, 70% of the data set is used for testing, and 30% is used for validation. It goes beyond evaluating the "fit" of the model to historical data toward the examination of the forecast errors when estimating fresh new data.

Forecast Model Deployment. Model deployment [21] involves getting the model and the resulting forecast in use by a customer. The user must participate to know how to exploit the model and decide how to visualise results. Monitoring

forecast model performances is a continuous process to ensure that the model deployed is still performing satisfactorily. Sometimes, models that performed very well in the past might deteriorate, leading to a more significant forecast error.

2.3 Discussion

This section introduced the state of the art for energy consumption and carbon footprinting environments. Big data analytics, machine learning and artificial intelligence are employed in the smart metering environment to extract functional patterns from the massive amount of data collected from the smart meters. Combining these techniques makes it possible to predict energy consumption and then estimate and predict carbon footprint. Measuring these consumption references and associating them with human behaviour and economic aspects (energy invoice) can encourage people to develop strategies to decrease their consumption. However, applying these analytics still faces numerous difficulties, as most utilities and customers are uncertain about the results produced by the analytics. Therefore, our work proposes an environment, namely GREENHOME, that enables the application of analytics and prediction combining different variables and models to observe energy consumption and CO_2 footprint in households under different perspectives.

3 SMART ENERGY AND CO_2 FOOTPRINT METERING ENVIRONMENT

GREENHOME is a smart metering energy and CO_2 footprint environment that analyses household energy consumption. GREENHOME lets scientists combine different variables to produce models that give different energy consumption perspectives to understand and compare. GREENHOME was implemented and validated through an experiment defined in the context of the Triple-A European project[5] willing to show homeowners that, with behaviour changes and investment in carbon-free technologies, both energy consumption and CO_2 footprint can decrease.

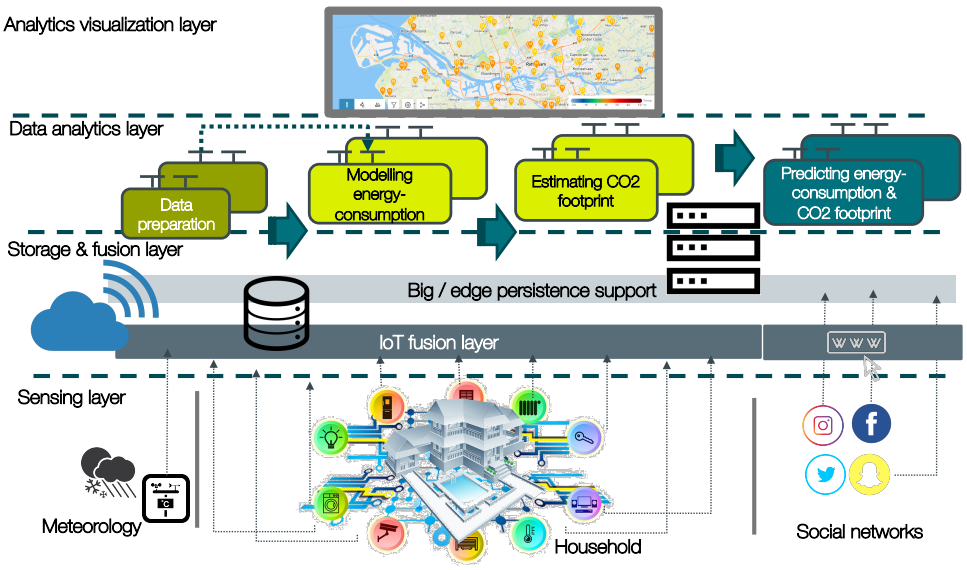


Fig. 1. GREENHOME general architecture

[5] http://www.triple-a-interreg.eu

The GREENHOME architecture is organised in four layers (see Figure 1). Two external layers are devoted to input and output. From the input side, GREENHOME is fed with data stemming from smart metering sensors inside and outside households, from meteorological services for exogenous variables, and we also consider a social aspect with data produced by social network services that communicate information about residents habits[6]. The output information of GREENHOME is given with visual results combining different dashboards. The visualisation layer is an interface to visualise different components using dashes. The primary purpose is to provide a decision-making tool using dashes, including tables, graphics, graphs, and other visual elements that best help understand data.

Two internal layers of GREENHOME are the data storage and fusion layers. They are externalised towards the cloud. GREENHOME relies on services devoted to IoT* data fusion, and REST* services provision for Web and storage services that ensure data persistence and data feeding to processes that run energy models.

The core of GREENHOME is the Data Analytics layer, which provides libraries for designing data science pipelines to build a different analysis method of a given problem (e.g., energy and CO_2 metering and consumption prediction).

3.1 Sensing, Fusion and Storage Layers

The sensing layer gives access to different metering tools (i.e., things) used to collect data from three types of sources: (i) social networks, which are REST services providing Tweets, Facebook posts, etc., producing insights related to energy consumption, (ii) weather stations in the specific locations, and (iii) sensors equipped near and inside the household to collect meteorology data. Combining a set of sensors, social data through a communication network can lead to different estimations of households' energy consumption and the CO_2 footprint. Data collected from things (sensors) connected to the Internet are sent to the cloud via communication protocols provided by the IoT fusion layer.

The IoT* fusion layer integrates heterogeneous data to produce consistent and useful collections. The edge persistence support provides communication between heterogeneous data from different sources and the data analytics layer. Sensor data fusion is performed using several algorithms, including central limit theorem, Kalman filter, Bayesian networks, Dempster-Shafer, and convolutional neural network [9, 35, 39]. The storage layer stores integrated data relying upon a combination of several systems such as HDFS and NoSQL systems like Apache HBase[7].

3.2 Analytics and Prediction Layer

The data analytics layer is the core of the metering environment. It provides analytics tools that implement different algorithms to prepare data, model energy consumption, estimate CO_2 footprint and predicts energy consumption.

Data Fusion Services. Stored data undergo two processes before being analysed: cleaning and integration. Data cleaning validates and pre-processes data integrating different sources into a dataset that can be analysed. Data cleaning consists of three phases: (i) adding metadata to the original data to document the procedure of data acquisition considering information related to a data source and the version of the collector; (ii) detecting bad data for tracking anomalous values and tagging them as missing or bad data; (iii) extracting features and deriving new data from raw datasets. Dataset integration merges different datasets and provides homogeneous datasets adapted for target analytics.

Data Preparation Services. These services transform integrated data to match the format expected by the data analytics services. Transformations include grouping or joining data. Depending on the purpose of the study, it is possible to prepare small datasets (i.e., samples) derived from an extensive initial dataset, applying traditional statistics.

[6]In the current GREENHOME version, the social data harvesting is not yet implemented and exploited. This concerns our current work.
[7]https://hbase.apache.org

Analytics and prediction services implement statistical and machine learning methods to estimate and forecast energy consumption and CO_2 footprint. Predictive techniques are based on models to explain, cluster, forecast the variables under study. The main output is trained models that predict the CO_2 footprint and energy consumption.

The visualisation layer uses the results to create graphics representing the relationship among variables, such as energy consumption and CO_2 footprint.

The study, modelling, estimation and prediction of energy consumption in households requires considering data concerning architectural, social, behavioural, technical, natural variables. These data are collected in different conditions and are not always available or cannot always be correlated because of mathematical or privacy constraints. Analytics environments must be flexible because they should provide insight into energy consumption with the data they have and then easily enhance their models when other datasets are liberated. The originality of GREENHOME is that new data providers, datasets and models can be added as new components that can be then used to produce, discover and predict consumption models and CO_2 footprint.[reviewer]

To validate our GREENHOME, we developed experiments based on a use case described next.

3.3 Experiments

We used as an experiment scenario the Triple-A project that aims to identify and describe the household energy consumption for increasing energy efficiency and reducing CO_2 emission of single-family houses[8]. The implemented use case targets the observation of a household in Picardie under the supervision of SPEE [28], an integrated service of energy renovation of private housing. SPEE uses smart meters to accomplish real-time measurements of the energy used for heating and specific electricity. The house understudy is a working-class house with red bricks built in 1926. A living space area of 85 m^2 with only gas as heating energy. Gas is used too for heating water. The living room is oriented southeast, and because of retirement, the single occupant of the house is all day all night at home. The indoor temperature, as programmed, is 20°C day and 17°C at night. Data collected are: (i) electric consumption, (ii) gas consumption, (iii) indoor/outdoor temperature, and (iv) outdoor humidity.

Other meteorological historical data were downloaded from the Meteoblue website[9]: (i) total precipitation, (ii) snowfall amount, (iii) total cloud cover, (iv) sunshine duration, (v) shortwave radiation, and (vi) wind speed and direction.

Electric and gas meters were built adapted to the house characteristics. A weather sensor was placed outside and protected from sunlight on the north facade to capture outdoor temperature and humidity. A comfort sensor was installed in the house where there is not much temperature and humidity variation. Sensed data are provided in 20 CSV files, collected between January 2018 and February 2019, with 4 CSV files containing two months observations.

Data includes energy consumption and gas arranged in a cumulative order; indoor/outdoor temperature and indoor/outdoor humidity. The gas consumption meter was installed on the 2nd of August. Gas consumption was excluded from the modelling hereafter to ensure credibility in the analysis. Note that data are timestamped[10]. Python 3.7 was used as a programming language[11]. The Python implemented application runs on a Docker [44] environment.

Methodology. The use case requirement was to estimate the electric end-use efficiency profile in buildings and carbon footprint to derive a decision support tool for the electric sector. According to [2], understanding the residential building

[8] https://github.com/javieraespinosa/Triple-A-household-energy-analysis
[9] https://www.meteoblue.com/en
[10] A dashboard providing a graphical view of the data is available at https://triple-a-demo.herokuapp.com
[11] It is an interpreted, functional, high-level programming language with dynamic semantics.

energy consumption as an independent statistical object is adequate for systematically accumulating the underlying data for residential building energy consumption and understanding the conditions of its energy consumption. Therefore, we proposed an energy consumption statistical system and explored effective statistical methods for studying building energy consumption.

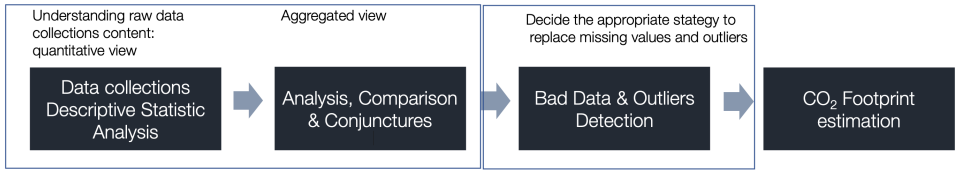


Fig. 2. Analytics metering pipeline

We designed an analytics metering pipeline encompassing four steps (see Figure 2): (i) data collections statistic characterisation implementing statistics and plots to discover trends and patterns; (ii) analysis, comparison and conjectures; (iii) data preparation* for deciding the appropriate strategy to replace missing values and outliers; (iv) modelling energy consumption and estimating the derived CO_2 footprint.

Fig. 3. Forecasting energy consumption pipeline

Then we designed a forecasting energy consumption pipeline consisting of three steps (see Figure 3): (i) computing a forecasting baseline using naive forest persistence; (ii) forecasting energy consumption without exogenous variables using ARIMA; (iii) forecasting energy consumption with exogenous variables using ARX. The implementation of these pipelines is described in sections 4, 5 and 6.

4 DATA COLLECTIONS PREPARATION

Preparing* data collections implies detecting and replacing outliers*. We applied three methods in the experiment: extreme value analysis (EVA), proximity, and projection. We show that the box plot provided in the extreme value analysis produces the best observation for outliers, and it was the one used for replacement in the experiment.

4.1 Quantitative Profile of Data Collections

A quick statistical information on the numeric column related to energy consumption per hour using the Pandas method `pd.describe()` shows the statistical description in Table 1.

The distribution of the values of the CSV is given as follows: (i) cumulative energy consumption with a timestamp and no specific time difference; (ii) cumulative gas consumption with timestamp and no specific time difference; (iii) external temperature and external humidity with timestamp recorded every 10 minutes; (iv) internal temperature and internal humidity with timestamp recorded every 10 minutes.

Table 1. Statistical description of the initial dataset

Count	**102**
Count	10152.000000
Mean	164.634161
Std	187.285460
Min	0.000000
25%	62.000000
50%	99.000000
75%	180.250000
Max	1985.000000

Note that there are missing observations after having a minimum equal to zero. To count missing data, we assigned a *true* mark to all values in the subset of the Pandas DataFrame that have zero values. Then we count the number of *true* values in each column. There were 148 values missing values in the electric consumption data, which is equal to 6 days. In contrast, only 15 observations were missing in the external temperature. The analysis did not retain the gas consumption dataset because it contained too much missing data. In the context of the project, the gas metering was an issue because we could not identify the right technology to use considering the characteristics of the observed household and the national regulations. This issue had implications for the reliability of the estimations. However, it was more important to avoid biased and wrong readings that could perturb the study.

We observed an enormous gap between the mean value (approximately 165) and the maximum value (approximately 1985). This issue required detecting outliers and replacing them[12]. The strategy here was to identify and analyse representative sample data. Therefore, cumulative electric consumption data was shifted toward its initial format. After computing the first discrete difference of each element in the consumption dataset, and due to missing slots, six values were found as huge negative numbers and were replaced by zero. The new values are the estimated actual consumption values grouped by timestamp (see Figure 4).

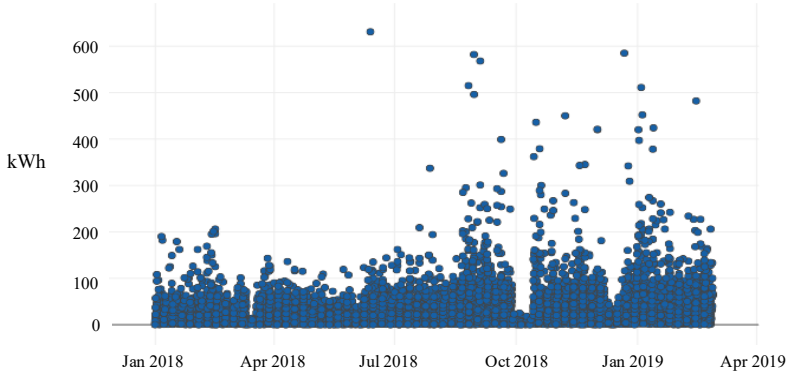
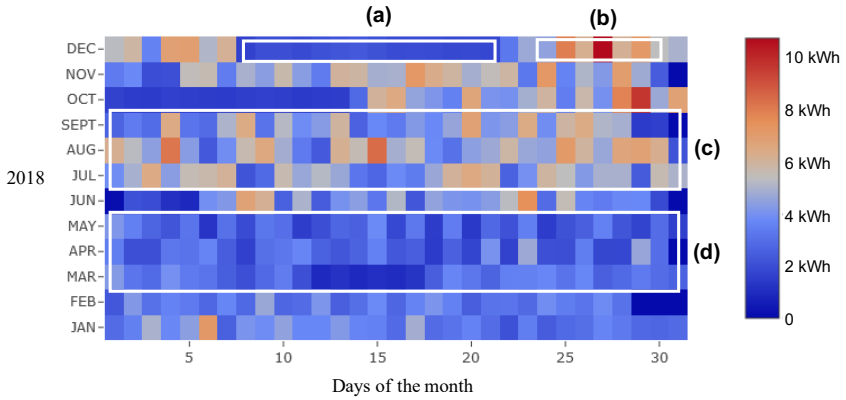


Fig. 4. Electric consumption before sampling between Jan. 2018 and Feb. 2019

[12]Different methods can be applied for replacing missing values. (i) Using a constant value that has meaning within the domain. (ii) Choosing a value from other randomly selected records. (iii) Estimating a value using a model. (iv) Computing the mean, mode or median of the initial set.

It was necessary to resample from the original data to create datasets, from which the variability of the quantiles of interest could be assessed without long-winded and error-prone analytical calculations [23]. In our experiment, data were sampled in two ways. The first sampling was done on the entire dataset as an hourly sampling. The main reason is that exogenous variables were used in the model, including temperature, for example, and it was not appropriate to use a unique temperature value for the whole day.



(a) Nerly zero consumption. Absence of the occupant due to holidays?
(b) Consumption peak. Visitors present for holidays celebration?
(c) Consumption increase. Due to cooling devices and high temperatures?
(d) Low consumption. Due to daylight-saving time?

Fig. 5. Daily energy consumption in 2018 and conjectures

Figure 5 shows a graph of the energy consumption after being daily re-sampled with some conjectures about this initial analysis. A long-range period in December was detected as nearly zero daily consumption. This absence of consumption can be explained assuming the absence of the occupant in the holidays period (see (a) in Figure 5). Close to the beginning of the year, there is a consumption peak. This peak might reflect the presence of other occupants, for instance, for celebrating holidays (see (b) in Figure 5). Spring months do not show high consumption. Particularly March, April and May have low daily consumption. This low daily consumption can be due to the increase in daylight saving (see (c) in Figure 5). In Summer, the fact that people need some cooling devices due to high temperatures increases the daily energy consumption in houses as shown in number (d) of Figure 5. Note that both dark red colours are considered outliers. In the project, we intended to validate these conjectures by analysing social media posts that could give insight on whether the inhabitant of the house was on vacation or in which periods the person was at home or not. This analysis introduces privacy issues that are difficult to address without careful processes. The house's inhabitant validated these conjectures, so our future work will develop human-in-the-loop techniques for completing and validating analysis results.

4.2 Extreme Value Analysis

Extreme value analysis (EVA) deals with the extreme deviations from the median of probability distributions [41]. A common way of approaching an extreme value problem is to divide the data into subsamples, then one of the extreme value distributions is fitted to those observations [45].

Outliers are often easy to spot in histograms. Indeed, the histogram in Figure 6 shows the presence of outliers. The histogram in Figure 6 divides the range of values into 12 groups based on the month and then shows the frequency — how many times the data falls into each group — through a bar graph.

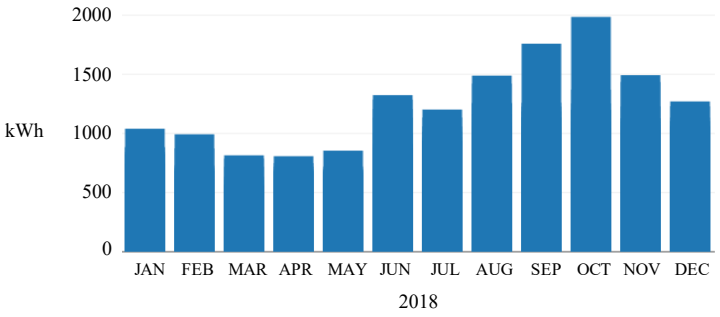


Fig. 6. Energy consumption per hour grouped by month

No outliers are detected in the sample. The recommended next steps are to plot a scatter plot of the data and a boxplot to observe outliers. Another plot that has been used is the scatter plot in Figure 7 (left) that groups data by month. In this case, values far from the group of the same month are considered outliers. Our graph shows outliers in June, August, September, October and November.

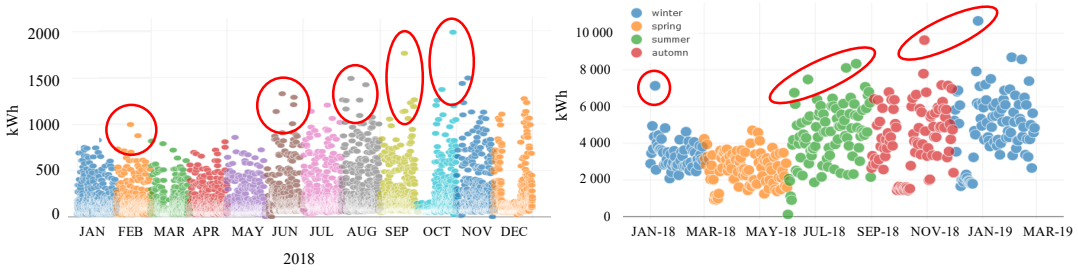


Fig. 7. Scatter plot for energy consumption per hour grouped by months (left) and daily energy consumption over one year (right)

A scatter plot after grouping data by seasons provides insight into the problem, including detecting outliers and analysing the change of behaviour over the seasons (see Figure 7 right).

The boxplot graph in Figure 8 spots outliers depicting groups of numerical data through their quartiles. It captured the summary of the data with a simple box and eased comparison across groups[13].

Observe that the median differs from one month and the other, with July having the highest median, May having the lowest consumption variation, and December having the highest consumption variation. Some data points not

[13]The function `boxplot()` of Pandas has been used to plot a boxplot. We also used the `seaborn` library from the Pandas library to generate the graph shown in Figure 8.

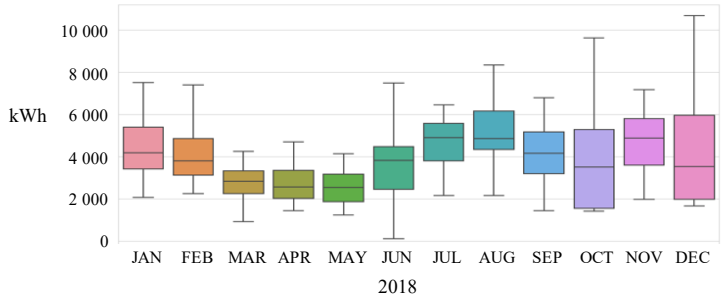


Fig. 8. Boxplot for daily energy consumption grouped by months

included between the whiskers were plotted as an outlier with a star (above 8000) in February. The graph shows the consumption per day for each month. It can be interesting to observe consumption per hour.

Therefore, another graph was plotted to spot outliers in each hour per month (see Figure 9). The mean value is the same somehow in all months, and numerous outliers are spotted for all months. The suite of data exploration tasks at different granularities with the applied methods showed the importance of the kind of strategy adopted for thoroughly understanding data from different perspectives. Again having a tool like GREENHOME that provides an environment that promotes this multifaceted exploration has been vital for the experiments and the project in general. Such multi-facet analysis can help answer questions about the sensors used for observing variables, about the habits of inhabitants, etc.

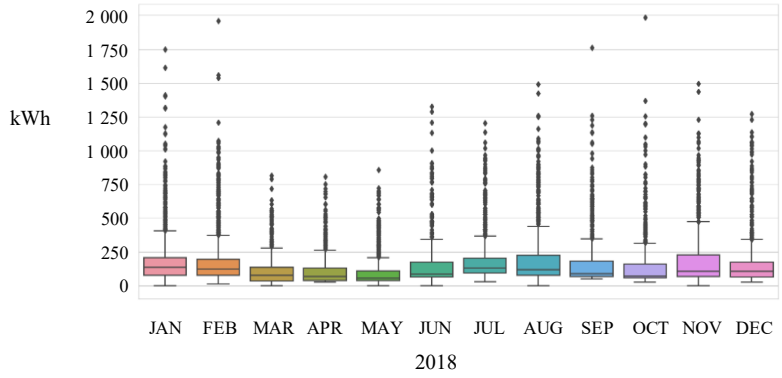


Fig. 9. Boxplot for hourly energy consumption grouped by month

4.3 Proximity Method

Given a dataset spread in a space, the measured distance between two data points in the dataset can be used to quantify the similarity between two data points. Consequently, data points being far from each other can be considered as outliers*. The proximity method assumes that the proximity of an outlier to its nearest neighbours significantly deviates from the proximity of the data point to most of the other data points in the data set [11].

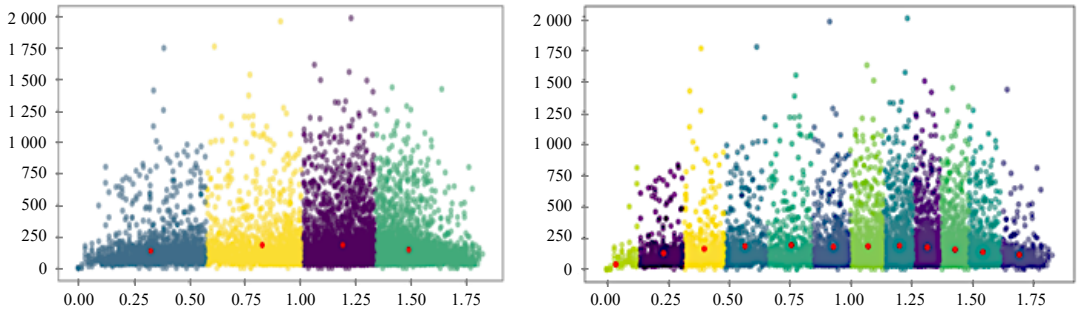


Fig. 10. K-means clustering for hourly electric consumption with k=4,12

K-means clustering is a proximity method [42] that partitions data into k groups assigning them to the closest cluster centroid. Once these centroids have been assigned, the distance between each object and a cluster centroid is calculated, those with the most significant distance are considered as outliers[14]. We defined four clusters assuming that there are four seasons in the year, and then we defined twelve, given the twelve months of the year. The clustering algorithm that clusters the dataset by month has a similar behaviour related to energy consumption resulting in a stable consumption for the whole month (see Figure 10 K-means clustering for hourly electric consumption with k=4,12). Clustering provides a view of the readings and understand whether readings and energy consumption are seasonal. Seasonality is an initial hypothesis regarding energy consumption.

4.4 Interquartile Range Method

Projection methods are relatively simple to apply and quickly highlight outliers [63]. We used the Interquartile Range Method (IQR) because it is well-adapted for exploring data with non-Gaussian distribution as in our experimental case. The IQR is derived from the difference between the 75th and 25th percentiles of the data. It identifies outliers by defining limits on the sample values that are a value of k above the 75th or below the 25th. K is defined as 3 or above to find extreme outliers[15].

```
Percentiles: 25th = 62.000, 75th = 180.250, IQR = 118.250
Identified outliers: 1050
Non-outlier observations: 9102
```

Then the IQR can be defined as the difference between the 75th and 25th percentiles already calculated. The cutoff of outliers was calculated as 1,5 times IQR. This cutoff was subtracted from the 25th percentile and added to the 75th percentile to give the definite limits of data. After running the above strategy, the following results were derived: 1050 values were detected as outliers as they lay below the 25th percentile equal to 62, or they rise above the 75th percentile equal to 180. As a result, the outliers represent 10% of the dataset.

[14] From `sklearn.cluster` library in Python, the K-Means function was used to cluster data classifying them into four groups of equal variance.
[15] The percentiles of the data series related to energy consumption were calculated using `percentile()` NumPy method that uses as parameter the data set and the percentile desired.

4.5 Comparison and Bad Data Replacement

A bad data item* is an outlier* that seems an unlikely observation produced when observing human behaviour. Three methods were used in the above technical experiment.

(i) The Extreme value analysis plotted by a histogram giving the first glimpse for discovering outliers, then a scatter plot to detect outliers easier. The basic plot drawn was the Box and Whisker plot to identify outliers.
(ii) The K-Means clustering algorithm was used to identify proximity between data points. The observation with high proximity to the cluster centre was considered as an outlier.
(iii) The mathematical approach IQR computed the series, out of which an outlier is identified. Based on the applied clustering method, about 1050 values were detected as outliers and replaced by the mean value according to each month.

The first quantitative exploration of the data content helped technicians in the project gain insight into how hardware (i.e., sensors) were working and the "quality of the readings". Energy consumption analysts could start having a first view of the energy consumption behaviour, mainly when data were organised into seasons. This shows the importance of performing comparative and multi-perspective datasets exploration.

5 COMPUTING HOUSEHOLD ENERGY CONSUMPTION AND CO_2 FOOTPRINT MODELS

A sound understanding of the determinants that drive household electricity consumption is needed for efficiently planning and analysing efficiency. We analysed variables to determine they influence energy consumption. We used the sensitivity analysis proposed by the Morris model for performing the study. Finally, given the computed electric consumption, we estimated the CO_2 footprint. The complementary question that we aim to answer is which variables within the household and external to it determine the energy consumption and contribute to increasing the size of the CO_2 footprint.

5.1 Sensitivity Analysis Using the Morris Model

Smart meters and home energy-monitoring services have produced data associated with variables that allow studying determinants of energy use and energy-related behaviours like the external temperature, external humidity, total precipitation, snowfall amount, total cloud cover, shortwave radiation, wind speed and wind direction.

We used the Morris Model to perform a sensitivity analysis to determine their influence on energy consumption. The Sensitivity analysis ranks inputs according to their influence on "energy consumption" output variability; that is, it screens out inputs, which have little or no influence on energy consumption. The results justify the choice of input values to calibrate the model used for forecasting energy consumption.

The Morris analysis specifying the percentage influence of each parameter on the output energy consumption is shown in Figure 11. Note that external humidity has the highest impact on the overall energy consumption, whereas the snowfall amount is trivial. The shortwave radiation also influences the energy consumption in the house, mainly due to lighting. A high value of total cloud cover means no radiation is exposed to the house, which requires using light, leading to an increase in energy consumption. The external temperature might be linked to an electrical device such as a fan that the occupant turns on when it is hot. For the time being and because we could not use social behaviour data, we have not performed a correlation study among the household energy consumption behaviour, the variation of exogenous variables and the actual actions of the inhabitants in a household.

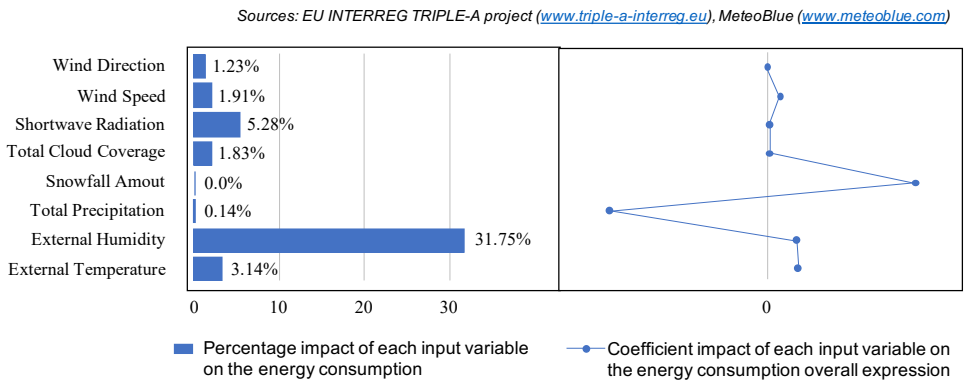


Fig. 11. Impact of different input variables on the energy consumption

Table 2. Carbon emission factor from the Triple-A project ($kgCO_2/KWh$)

Member state	Displaced electricity	Natural gas	Heating Oil	Biomass
UK	0.519	0.216	0.298	0.039
Netherlands	0.530	0.204	0.267	0/0.395
France	0.09	0.241	0.329	≈ 0.013
Belgium	0.258	0.202	0.279	0

5.2 Mathematical Estimation of the CO_2 Footprint

The carbon footprint is a measure of the total amount of Carbon Dioxide (CO_2) and other greenhouse gas emissions directly or indirectly caused by an activity or accumulated over the life span of a product, person, organisation or even a city or state [20]. A CO_2 footprint determines the emission of greenhouse gases produced due to, directly and indirectly, human activities. The methodologies for calculating the CO_2 footprint are still evolving even if the carbon footprint is becoming a standard criterion for managing greenhouse gas.

Each country uses different sources and input variables to model annual energy use for gas and electricity and derives an estimation of CO_2 emission. Figure 2 provides the breakdown by Triple-A partner countries. The calculation used to generate annual carbon savings for this project is given by:

$$\frac{tCO_2}{a} = \frac{[(EDPM(kWh) - EDPI(kWh)) \times REF(kgCO_2/kWh)]}{1000}$$

where *EDPM* is the Energy demand before the measure, the *EDPI* is the Energy demand post-installation, and the *REF* is the relevant emissions factor.

The formula used to calculate CO_2 emissions based on electrical consumption inside a house is as follows:

$$\frac{tCO_2}{a} = [\text{Energy consumption (kWh)}] \times \text{relevant emissions factor } (kgCO_2/\text{ kWh}) \text{ / } 1000$$

However, the choices of the Triple-A project about the carbon emission factors concerning each country may not be completely accurate. Thus, we decided to use values from the Réseau de Transport d'Électricité (RTE), which continuously provides an indicator of the carbon footprint of electricity generation in France, expressed in grams of CO_2 per kWh generated.

Figure 12 shows the estimation of CO_2 footprint for each day of the 2018 year in our use case. The same peaks spotted in electric consumption are spotted in the CO_2 footprint. These observations suggest that energy consumption

must be reduced to reduce the CO_2 footprint. Actions must be adopted, especially at Christmas and some months in summer. Changes in habits can eventually lead to other energy consumption behaviour that will be compared in other experiments in the Triple-A project, which gives context to our experiments.

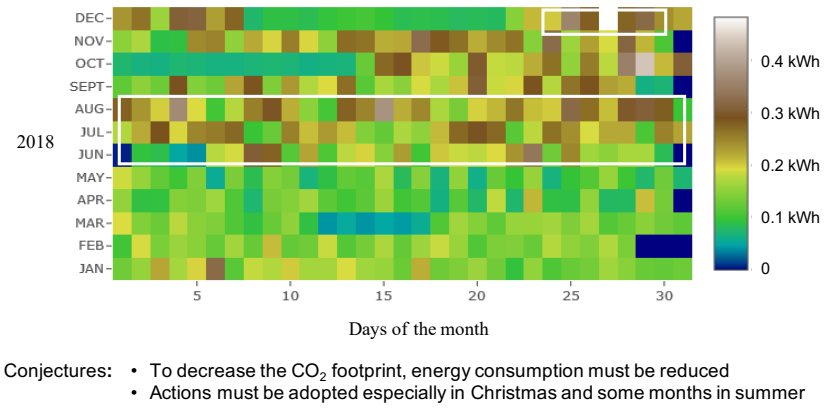


Fig. 12. Carbon footprint of hourly energy consumption

6 PREDICTING HOUSEHOLD ENERGY CONSUMPTION AND CO_2 FOOTPRINT

GREENHOME provides three energy forecast methods that use smart meters measurements and weather data to predict energy consumption in buildings. The pipeline implemented for studying energy consumption first applies the naive forecast model, the ARIMA model, and the ARX model with other inputs that might decrease the performance gap. These models provide different perspectives on energy consumption in a household. As said before, the GREENHOME environment promotes the comparison of these different perspectives. Beyond the final performance assessment of the models where one numerically performs better than the other, the variation of the criteria adopted for implementing the pipelines seems more revealing (e.g., choosing exogenous variables or no). The complementary results determine the type of questions that can be asked given specific combinations of variables.

6.1 Naive Forest Persistence Model

The naive forecast persistence model[16] consists of three steps: (i) preparing the dataset to create a lagged representation for each observation; (ii) using a resampling technique for splitting the dataset into train and test fragments; (iii) measure performance to evaluate the model. e.g., mean squared error. The pseudocode of the function and its complexity is given in Figure 13. For our experiments, the complexity of the algorithm did not determine the execution of the pipelines. However, when data collections increase volume, it is capital to consider this complexity for choosing the size of test, training and validation data sets concerning the available computing resources.

The persistence algorithm uses the value at time `t-1` to expect the predicted output at time `t`. The creation of a lagged representation of each observation means that given the record at `t-1`, the record at `t-1` is predicted. To fragment the

[16] https://www.sciencedirect.com/topics/engineering/persistence-model

```
Input: D as a dataset containing columns related to a set of
energy_consumption observation over a period of time (t).
1: For each observation(t) in the D
2:     Create a lagged representation observation(t-1)
3: Separate dataset into train(70%) and test data set(30%)
4: Perform a persistence algorithm
5:     model_persistence(x) = x
6: Evaluate the model
7:     Use walk-forward validation method
8:             For x in test_x
9:                   pr = model_persistence(x)
10:                  predictions.append(pr)
11:    Calculate mean squared error.
Output: predictions of the end 30% period of t with a mean squared error
for evaluation between predicted consumption and real one.
```

```
Complexity:
it requires O(1) space for every modification: just store the new data. Each
modification takes O(1) additional time to store the modification at the end
of the modification history.
```

Fig. 13. Persistence model and complexity measures persistence

dataset into training and test datasets, we made a classification of 99% for training and 1% for testing given the small size of the initial dataset. The persistence method can be defined as a function that returns the input provided.

The persistence model was evaluated on the test dataset using the walk-forward validation method[17]. The Walk-forward validation is a method where the model predicts each record in the dataset one at a time. Predictions were made for each record in the test dataset. The predictions were compared to the actual values. The computed residual mean squared error was `RMSE=77.835`.

The plot in Figure 14 "Persistence forecast model" shows the training dataset and the diverging of the predicted line from the actual values. Note that the model is a step behind the initial values. The graph is not stationary and varies a lot, which limits the persistence model.

The naive forecast persistence model is a baseline for the forecast problem; that is, if any other forecast model achieves a performance at or below the baseline, the technique needs to be improved or abounded[18].

6.2 ARIMA Model

An ARIMA model[19] [26] was designed and developed to solve the forecasting problem of household energy consumption (see Figure 15).

The model was configured both manually and automatically. Once the ARIMA model was used, its residual error was calculated. The standard notation is `ARIMA(p,d,q)` where `p` denotes the number of lag records included in the

[17] https://en.wikipedia.org/wiki/Walk_forward_optimization
[18] This is essential in the forecast problem because it gives an idea about how well all other models perform on the problem.
[19] ARIMA is a class of statistical model for analyzing and forecasting time series data.

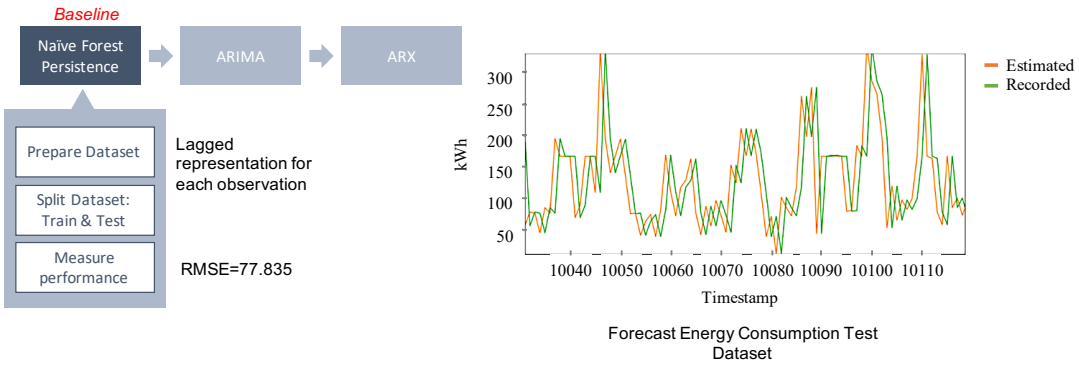


Fig. 14. Persistence forecast model

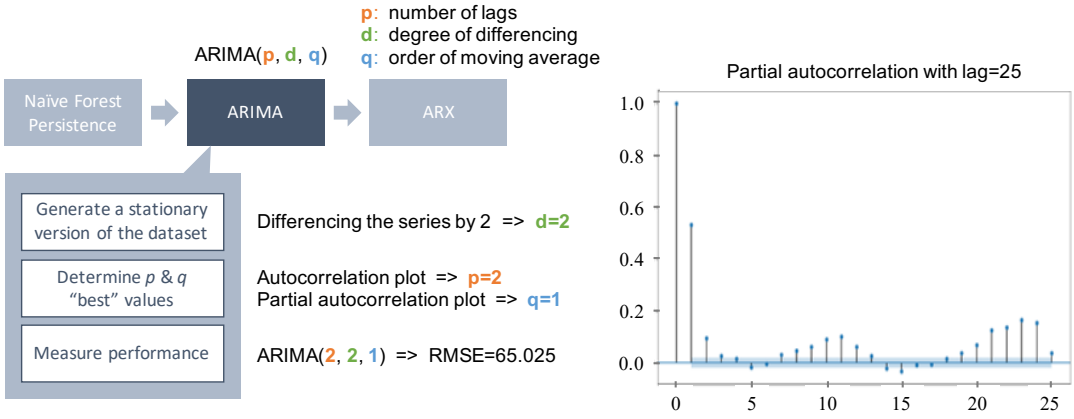


Fig. 15. ARIMA model pipeline

model, d denotes the degree of differencing, (i.e. the number of times the raw records are differenced) and q denotes the order of moving average, which is the size of moving window.

The pseudocode and complexity measures of the ARIMA model are shown in Figure 16[20].

Regarding complexity, ARIMA requires O(1) space for every modification to store the new data. Each modification takes O(1) additional time to store the modification at the end of the modification history. Components for measuring complexity are as follows:

- `Ms` = person hours of setting up the data and computer program for parameter estimation.
- `C` = computer use costs of analysis
- `MT` = mans hour of interpreting and tabulating computer results.
- `T` = number of observations in the dataset.

[20]Note that wherever q appears, it is multiplied by a factor of 2. This is to incorporate the fact that moving average and mixed processes are more complicated than an auto-regressive process.

```
Input: D as a dataset containing columns related to a set of
energy_consumption observation over a period of time.
1: Plot D
2: While (graph is non_stationary = True)
3:     smoothen graph to make it stationary
5:     Find best level of differencing
6: d = best level of differencing
4: Plot ACF/PACF(Stationary_graph)
5: If ACF shows no lags
6:     p = 0
7: else
8:     p = nbre of lags
5: If PACF shows no lags
6:     q = 0
7: else
8:     q = nbre of lags
9: Grid search ARIMA hyperparameters
10: for p: 0 to 4
11:    for d: 0 to 2
12:          for q: 0 to 4
13:                model = fit_model_ARIMA(p,d,q)
14:                mse = evaluate_arima_model()
15:                if mse < mse_initial
16:                      opt_model = model
Output: ARIMA_model(p,d,q)
```

Fig. 16. ARIMA model and complexity measures

- (p,d,q) = vector with components p, order of the AR portion of the model; d, degree of differencing to achieve stationary; and q order of the MA portion of the model.
- Ms, C, and MT are each positively related to T, p, d, and q.

For simplicity, complexity costs will be measured as:

$$\begin{aligned} Ci &= h.Ms + C + h.Mt \\ &= h.Ms(T,p,d,q) + C(T,p,d,q) + h.MT(T,p,d,q) \end{aligned}$$

Where h is the wage rate of the investigator.

(1) $Ms \sim \beta_1 + \beta_2.T + \beta_3.(p+2q+1)$

(2) $C \sim \beta_4.T + \beta_5.(p+2q) + \beta_6.T.(p+2q+1) + \beta_7.T.(p+2q+1)^2$

(3) $MT \sim \beta_8.(p+2q+1)$

For Ms, β_1 reflects the time needed to write standard subroutines for the numerical computation of parameters of the marginal distribution, the time needed to write a program section transition from the original dataset to the d^{th} differences, and program debugging time. β_2.T time needed to tabulate and check data. $\beta_3.(p+2q+1)$ measures the time needed to write p+q+1 integration routines.

For C, β_4.T measure costs of compiling data, printing predictions, and computing statistics. $\beta_5.(p+2q)$ reflects the cost of compiling the remainder of the program deck. $\beta_6.T.(p+2q+1)$ measures costs of computing predictions and obtaining plots of p+q+1 marginal. $\beta_7.T.(p+2q+1)^2$ measures cost of performing p+q+1 integrations. For MT, $\beta_8.(p+2q+1)$reflects the time needed to read and interpret the results of the analysis.

Manual Configuration. `ARIMA(p,d,q)` requires the parameters `p`, `d` and `q`. Usually, the configuration is done manually[21]. As shown in Figure 17, the energy consumption data series is not-stationary.

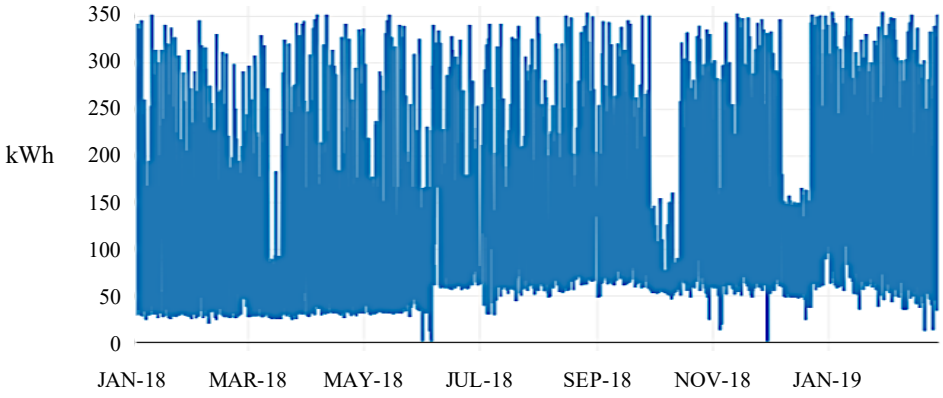


Fig. 17. Non-stationary plot of energy consumption per hour

A stationary version of the series is derived after differencing the original series, followed by a stationary test of the new data series. The unit root is done as a test for stationary. Unit root if found, then the time series is not stationary. The augmented Dickey-Fuller test gives a test statistic value -25.2, smaller than the critical value at 1%, equal to -3.43. This accepts the rejection of the null hypothesis at a high level. The idea of rejecting the hypothesis confirms that the process has no unit root, and therefore the series is stationary.

The difference between the test statistic value and the critical value is more than 20. Therefore, any value of d greater than 0 could be considered. This means that a differencing level of 2 can be used, and so `d=2`. Now, both the lag values and the moving average parameters, `p` and `q` should be selected. These values can be derived from the autocorrelation function plots and the partial autocorrelation function plots. By default, all lag values are plotted, which is a noisy plot. This requires a good lag value definition, and as it is an hourly prediction, the best-chosen lag value is 25 since a similar consumption pattern happens at the same hour of the previous day.

The left side of Figure 18 shows auto-correlation graphs. The first graph is condensed, with nothing to visualise. Therefore, as mentioned, a lag=25 is significant to the plot. The 2nd plot shows a correlation of 0.6 at lag=1, 0.4 at lag=2, and then 0.2 at lag=3. It is straightforward to see that lag=2 results in a good starting of p at 2.

The right side of Figure 18 presents the partial auto-correlation graph with lag=25, indicating a good starting value for q=1. The graph shows a partial auto-correlation equal to 0.55 at lag=1, then it drops significantly to 0.1 at lag=2, then there is no correlation at `t-2`. The best value for `q=1`. This analysis suggests a start with `ARIMA(2,1,2)` that gives an `RMSE=65.025` which is quite lower than the value generated by the persistence model.

Configuring ARIMA using Grid Search. To confirm the manual results, a grid search can be done to find best ARIMA parameters to ensure that no other combination can result in better RMSE performance. The search will skip values that will not converge. The values to search are: p: 0 to 4, d: 0 to 2, q: 0 to 4. This implied 300 runs of test harness, it took one hour to execute[22]. The results shows that the best ARIMA model is `ARIMA(2,1,3)` with an `RMSE=64.043`, used next.

[21] The method `statsmodels.tsa.stattools.adfuller()` was used as a unit root to verify if data is stationary.
[22] Here, the function `ARIMA()` provided in the `statsmodels.tsa.arima_model` library, and the function `mean_squared_error()` provided by the `sklearn.metrics` library were used.

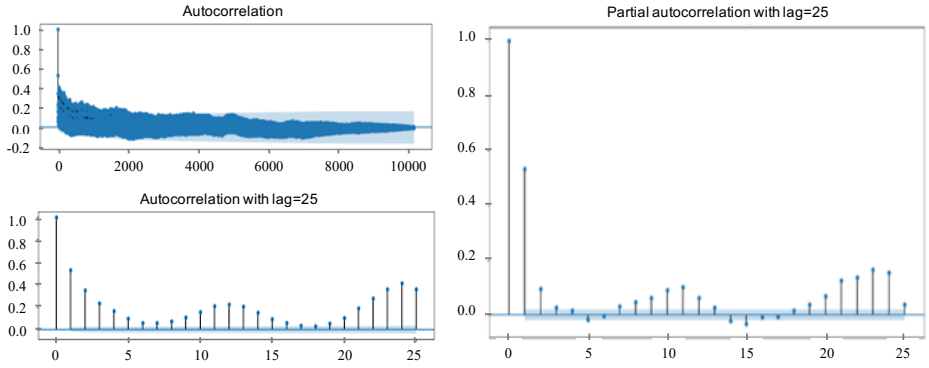


Fig. 18. Auto-correlation plot (left) and partial auto-correlation plot (right)

Table 3. Brief statistics for residual error

Count	102
Mean	-0.348572
Std	64.358404
Min	-153.106851
25%	-44.182336
50%	-4.695801
75%	28.659611
Max	241.561941

Review Residual Error. As a final validation of the chosen model, a review of the residual error forecast should be done. As an ideal case, the distribution of errors has to be a Gaussian distribution with the mean equal to zero. Brief statistics and plots can check this.

The mean is a non-zero value of -0.5. This value assures that the predictions are biased. The distribution of residual error is shown in Table 3.

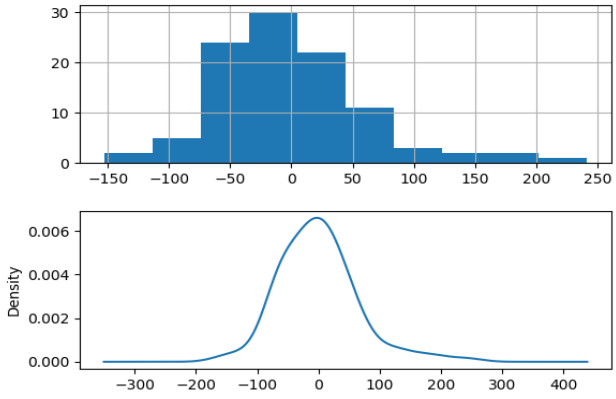


Fig. 19. Distribution of residual error

The plot suggests a Gaussian-like distribution with a long middle tail. This information can be used to bias the correct prediction by adding -0.589739 to each forecast made. The predictions performance changed from 64.043 to 64.042. Therefore, this bias correction can somehow be ignored, considering that the bias correction will increase complexity and cost without even improving performance in the study case because the performance did not change at all.

Model Validation. The selected model must be validated. The final RMSE value is 64.121, which is not too far from the previous calculated and expected value of 64.043. Figure 20 shows a plot of each prediction and expected value for the time steps in the validation dataset. Some observations have (almost) the same values as the predicted ones whenever there was no significant deviation between one hour and the other. The model has a significant performance gap whenever there is a sudden change in the hourly energy consumption. Adding exogenous parameters as input to the forecast model can be improved to improve the forecast and reduce the performance gap, which is explained in the next section.

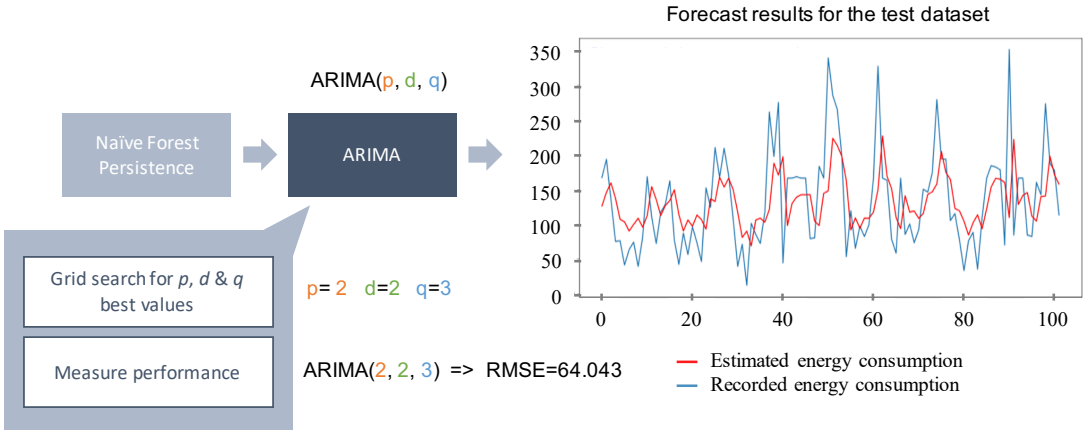


Fig. 20. ARIMA forecast model plot

6.3 ARX Model

The ARX model is an autoregressive method with exogenous inputs (independent of the process to model)[23]. Autoregressive models express a univariate time series y_n as a linear combination of past observations y_{n-1} and white noise V_n and are mathematically expressed as [37]:

$$y_n = \sum_{i=1}^{m} (a_i . y_{(n-i)} + v_n)$$

Where a_i and m represent respectively the auto-regressive coefficient and auto-regressive order. Considering inputs r_n and output S_n, the ARX model can be mathematically expressed as proposed in [37]:

$$s_n = \sum j = 1m(a_j . s_{(n-j)+\sum j=1m(b_j . r_{n-j}+u_n)})$$

Where u_n is white noise and a_j and b_j are p*p and p*q matrices, respectively.

ARX can be practical and effective when the parameter to be estimated a linear correlation with the input parameters of the algorithm. It is also effective for determining the order of the system. Thus, it is necessary to evaluate the order

[23] According to Diversi et al. [24], ARX is the simplest model within the equation error family. It has many practical advantages concerning estimations, and predictive use since its optimal predictors are always stable.

of the ARX polynomial to determine the order of the polynomial that results in the least cost and error. The AKAIKE criterion can be used to determine the most suitable order of the system [4]. The AKAIKE criterion is defined as an estimator of the relative quality of statistical models for a given set of data.

Forecasting Energy Consumption with ARX.. The external temperature, external humidity, wind direction, and total cloud coverage were the exogenous variables given as input for the forecast model of our experiment. The use of these variables required implementing an auto-regression model that considered the change of the electric consumption behaviour according to the exogenous variables. Different input data were used in the model, out of which the following lead to the minimum RMSE value with the highest performance about the AKAIKE value explained before. The linear regression expression went as follows, with `y: consumption, u0: external temperature, u1: external humidity u2: wind direction, u3: radiation, u4: wind speed`.

$$y[k] = +0.537388y[k-1] + 0.2407662u0[k] + 0.503764u1[k] + 0.010898u2[k] + 0.035441u3[k] + 0.105516u4[k] \quad (1)$$

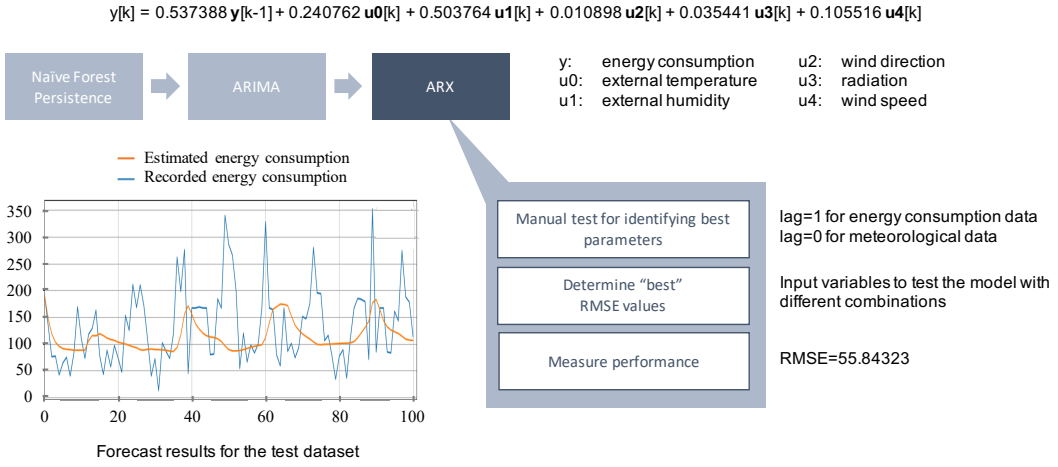


Fig. 21. ARX forecast model plot

Figure 21 ARX Forecast Model plot shows that the model forecasts a month and a half period and a high-performance gap with high peaks. Despite these issues, the forecast model successfully worked, as shown hereafter.

For the RMSE, a manual check for best parameters using different input variables was done to test the model with different combinations. The combination mentioned above of input variables led to the lowest RMSE=55.843230.

6.4 Discussing Results

As the results have proved, both models work as long as they produce an RMSE value less than the RMSE value of the baseline model, that is, the persistence model. However, the ARX model has an RMSE value less than the ARIMA model, and this is so logical because the ARIMA model does not consider exogenous variables rather than previous observations as an input for its forecast model. The idea of measuring energy consumption is of great interest and is full of magic and weirdness because energy consumption behaviour is unpredictable. However, there are some cases where predictions

can come true. The concept of predicting energy consumption without considering different indirect variables may work but not for every case. Much more, indirect variables that impact energy consumption differs from one condition to another. For example, the household in the case study uses electricity only for lighting, while other households use electricity for heating. For this reason, it seems nonsense to take into consideration the architecture of the house in the first case, but it can have a substantial impact in the second case. For our future study in the Triple-A project, the objective will be to produce datasets concerning a farm of houses with different architectural characteristics and locations. The interest in combining analysis with various patterns (architectural characteristics, exogenous variables determined by geographical location, habits depending on the number of inhabitants and socio-economic group, etc.) will lead to more accurate insight into energy consumption and carbon footprint.

7 SMART METERING SYSTEMS AND APPLICATIONS

A smart metering system* is an integrated infrastructure of smart meters, communication networks, and data management systems that enables two-way communication between utilities and customers [6]. The two primary functions are monitoring and control. Monitoring allows us to understand the way energy is consumed or generated at home and display historical data on demand. Control indicates if the energy management system can act on one element of the energy flow in a house or a building (e.g., switch on/off an appliance, adjust the in-house temperature, etc.). Combined with customer technologies, the objective of a smart metering system is to encourage customers to reduce energy consumption and carbon footprint [53]. It also allows utilities to offer incentives to customers to reduce peaks in energy demand and consumption at certain times. The following lines describe examples of these types of systems and compare them with our proposal.

7.1 Social Smart Metering

Understanding energy consumption behaviour is an essential element in sustainable studies. Energy consumption related information could be extracted from user-generated content posted on social media. Such work was proposed in [52], where a pipeline helps identify energy-related terms in Twitter posts. Twitter posts were classified into four categories related to dwelling, leisure, food, and mobility according to the activity performed. A web application was also developed that allows end-users to check their energy consumption based on analysis driven in the pipeline. The main thing that makes social media data trending is that traditional ways of getting data, including smart meters, are costly and may lack contextual information.

7.2 Netatmo Application

It is an easily configured application controlled by a smartphone (or tablet) to monitor and record the given local environment. Netatmo weather stations consist of several sensors, which monitor inside and outside air temperature (specified manufacturers accuracy: ± 0.3 °C) and relative humidity (± 3%), as well as indoor barometric pressure (± 1mb), carbon dioxide concentration and noise pollution. Optional additional measurements include precipitation and wind, although these modules are less frequently purchased and data are less available. Data is transmitted wireless, using a combination of Bluetooth and Wi-Fi, to the cloud where it can be accessed via a smart device, as well as being made available online via a "weather map" on the Netatmo website with observations updated every 5 min[24].

[24] Wikipedia, "Estimation Theory", 2019. [Online]. Available: https://en.wikipedia.org/wiki/Estimation_theory.

7.3 TOON Application

Toon is a smart thermostat solution developed by Quby (a company based in the Netherlands). The device offers a touch screen display through which users can set their preferences. Toon is not reduced to just thermal management. The device can provide valuable data on the building's energy consumption and be used for security purposes. Toon can interface with other smart devices such as smart plugs, Amazon Alexa and Philips Hue Lighting. Toon is an internet-connected device, thus allowing users remote access to change or update settings. Below is a summary of the specification of Toon.

Utilities such as Eneco (Dutch utility) are also able to collect anonymous and aggregated data. This data is helpful for scheduling, planning and allows end-users to compare their usage patterns to similar households. Subsequently, this data can be used to propose an optimised heating schedule [15].

7.4 Smart energy projects[reviewer]

EU-DEEP[25] focuses on the importance of distributed energy resources by addressing technical challenges, economic values and business models and drawing a set of recommendations. IntUBE[26] deals with the energy efficiency of single and groups of buildings. eDIANA[27] develops middleware and platforms to integrate buildings as nodes in the grid. A further issue concerns making users aware of energy and environmental issues. The project BeAware[28] proposes an interactive game that monitors the environment to advise users and award them according to their behaviour. The projects INTEGRAL[29] and SmartHouse/SmartGrid[30] propose multi-agent systems and home gateways to control local energy production, energy market, demand-side, and load forecast. The AIM project[31] models and manages domestic appliances.[reviewer].

7.5 Comparison

Existing systems vary a lot in the sensing technology used for sensing variables. This difference too introduces disparities in the type of observed variables. The system in Picardie in the Triple-A project was done in cooperation with QUARTUM that provides sensors for measuring/collecting electricity consumption ("electricity sensor"); measuring/collecting temperature and humidity outside ("weather sensor"); measuring/collecting temperature and humidity inside ("comfort sensor"); measuring/collecting gas consumption which is optional as not every household uses gas for heating ("gas sensor"). The system also uses a tablet for showing the collected data. The data collected by the electricity, the weather and the gas sensors are transmitted via radio frequency (433 MHz) to the tablet. The comfort sensor sends information to the tablet via Bluetooth.

In general, systems monitor real-time energy consumption for providing an aggregated visual view to household inhabitants. However, none of the studied environments addresses the carbon footprint of the consumed energy at the level of households and buildings. Of course, energy consumption is related to energy consumption invoices but has a less environmental impact. Only two environments worked on predictions of energy consumption under the aim of

[25] http://www.eudeep.com
[26] http://www.intube.eu
[27] http://www.artemis-ediana.eu
[28] http://www.energyawareness.eu/beaware
[29] http://integral-eu.com
[30] http://www.smarthouse-smartgrid.eu
[31] http://www.ict-aim.eu

decreasing consumption. Both systems need the use of solar panels. The idea of providing a new environment that predicts energy consumption while estimating carbon footprint is yet to come.

Prediction can be important for household inhabitants because they can plan and organise their activities beforehand and learn from previous behaviour patterns to adjust them to adopt a green behaviour. The economic dimension derived from a more intelligent energy consumption pattern can encourage the homeowners to invest in more accurate sensing material and even in house structural modifications that can help to reduce consumption and carbon footprint. The question is to determine to which extent human behaviour patterns can reduce these metrics and at which point inhabitants need to modify architectural and structural modifications of the house to achieve greener consumption.

8 CONCLUSION AND FUTURE WORK

We have proposed GREENHOME, a smart metering energy and CO_2 footprint environment, using a toolkit for modelling and predicting energy consumption in households at different granularities and from different perspectives. The experimental environment was implemented in a house in Picardie, where electrical consumption was predicted using three different models: the persistence model, ARIMA model, and ARX model after detecting anomalies to best fit the model on the given data. The carbon footprint was also estimated using some mathematical equations.

The results show that the implemented ARX model, which adds exogenous variables as an input, results in less RMSE value and better performance than the ARIMA model utilised. After a sensitivity analysis implemented using the Morris method, we included exogenous variables to check which variables impact the most on the hourly consumption. External temperature and external humidity were the two most significant variables that affect consumption.

By applying the different models for estimating energy consumption and forecasting, we identified the importance of combining technical data stemming from sensors installed in the household with meteorology, location, architectural, urban data, and social data representing inhabitants' actual habits and behaviour. The study of variables sensitivity and the experimentation of their use for forecasting energy consumption in households copes with the Triple-A project's objectives that are willing to increase awareness and easy access to the visualisation of variables that play a role in their energy consumption and carbon footprint. This strategy shall result in increased adoption of behaviour that can decrease energy consumption and carbon footprint. This strategy will also lead to the investment of house owners to improve their houses with low-carbon technologies. At the same time, tools like GREENHOME will have to evolve, proposing user-friendly decision-making interfaces for house owners and governmental decision-makers to identify how to promote low-carbon behaviours and develop policies including funding to achieve the development of green households.

Energy consumption and CO_2 emissions in the household is a crucial element to achieve sustainability from social, economic, and environmental perspectives. The sustainable development goals agreed in 2015 by the United Nations discuss the importance of addressing energy consumption and CO_2 emissions in many different domains. In [58] authors identify SDG-related research initiatives and activities and found that the majority of contributions to by the IEEE and ACM research communities have mainly focused on the technical aspects. At the same time, there is a lack of holistic social good perspectives. In particular, the problem of energy consumption in households calls for more holistic strategies. Considering holistic approaches is part of our current work.[reviewer]

Smart meter data analytics is a promising area that incorporates different fields of science, including the machine-learning field. It is, without doubt, a topic that will grow more important as long as the smart grid topic is developing, where all parties involved should reap the environmental and economic benefits of progressing load forecasting and estimating the carbon footprint.

Therefore, it is crucial to develop multi-facet analytics platforms to predict energy consumption in buildings and estimate their CO_2 footprint. This multifaceted view can contribute to producing a better understanding of the behaviour to adopt to help decrease global warming. For the time being, studies focus on smart metering produced data that can be more or less precise depending on the installation of smart meters in households. Other data issued from social media or annotated explicitly by the inhabitants can be complementary. However, integrating different data retrieved under different conditions and contexts calls for the combination of analytics models. This combination must follow rules that still need to be stated. As GREENHOME, platforms that serve as analytics labs where data scientists can combine and explore different models on top of heterogeneous datasets need to emerge and be consolidated.

ACKNOWLEDGMENTS

This work was partially funded by the SYREL team of the Grenoble Electrical Engineering Laboratory (G2Elab) and the Triple-A European Project (http://triple-a-interreg.eu).

REFERENCES

[1] Manar Amayri, Stephane Ploix, and Sanghamitra Bandyopadhyay. 2015. Estimating Occupancy in an Office Setting. (2015), 72–80.

[2] M. Arif, T. Brouard, and N Vincent. 2006. A fusion methodology based on Dempster-Shafer evidence theory for two biometric applications. In *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 590–593. https://doi.org/10.1109/ICPR.2006.68

[3] Rachad Atat, Lingjia Liu, Jinsong Wu, Guangyu Li, Chunxuan Ye, and Yi Yang. 2018. Big Data Meet Cyber-Physical Systems: A Panoramic Survey. *IEEE Access* 6 (2018), 73603–73636. https://doi.org/10.1109/ACCESS.2018.2878681

[4] R Baragona and F Battaglia. 2004. Projection Methods for Outlier Detection in Multivariate Time Series. *Sis-Statistica.It* 2000 (2004), 107–118. http://old.sis-statistica.org/files/pdf/atti/sessioneplenarie2006{_}107-118.pdf

[5] Armin Barghi, Amir Reza Kosari, Maede Shokri, and Samad Sheikhaei. 2014. Intelligent lighting control with LEDS for smart home. *Smart Grid Conference 2014, SGC 2014* (2014), 1–5. https://doi.org/10.1109/SGC.2014.7090861

[6] María del Carmen Bas, Josefina Ortiz, Luisa Ballesteros, and Sebastián Martorell. 2017. Evaluation of a multiple linear regression model and SARIMA model in forecasting 7 Be air concentrations. *Chemosphere* 177 (2017), 326–333. https://doi.org/10.1016/j.chemosphere.2017.03.029

[7] Jorg Breitung. 1994. Some simple tests of the moving-average unit root hypothesis. *Journal of Time Series Analysis* 15, 4 (jul 1994), 351–370. https://doi.org/10.1111/j.1467-9892.1994.tb00199.x

[8] E U Building and Stock Observatory. 2014. Energy Performance of Buildings Directive. *Structural Survey* 23, 1 (2014), 1–7. https://doi.org/10.1108/ss.2005.11023aab.001

[9] Lee Chapman, Cassandra Bell, and Simon Bell. 2017. Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. *International Journal of Climatology* 37, 9 (2017), 3597–3605. https://doi.org/10.1002/joc.4940

[10] Hamed Chitsaz, Hamid Shaker, Hamidreza Zareipour, David Wood, and Nima Amjady. 2015. Short-term electricity load forecasting of buildings in microgrids. *Energy and Buildings* 99 (2015), 50–60. https://doi.org/10.1016/j.enbuild.2015.04.011

[11] A. C. Davison and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

[12] Paraskevas Deligiannis, Stelios Koutroubinas, and George Koronias. 2019. Predicting Energy Consumption Through Machine Learning Using a Smart-Metering Architecture. *IEEE Potentials* 38, 2 (mar 2019), 29–34. https://doi.org/10.1109/MPOT.2018.2852564

[13] Narendra Kumar Dhar, Nishchal Kumar Verma, Laxmidhar Behera, and Mo M. Jamshidi. 2018. On an Integrated Approach to Networked Climate Control of a Smart Home. *IEEE Systems Journal* 12, 2 (2018), 1317–1328. https://doi.org/10.1109/JSYST.2016.2619366

[14] Home Display. 2011. Smart Meter / IHD. I (2011), 1–15.

[15] Roberto Diversi, Roberto Guidorzi, and Umberto Soverini. 2010. Identification of ARX and ARARX Models in the Presence of Input and Output Noises. *European Journal of Control* 16, 3 (2010), 242–255. https://doi.org/10.3166/ejc.16.242-255

[16] Dieter Elz. 2007. Bioenergy systems. *Quarterly Journal of International Agriculture* 46, 4 (2007), 325–332.

[17] Hassan Farhangi, S Rangan, and H Zhang. 2012. Smart grid and ICT's role in its evolution. In *Green Communications: Theoretical Fundamentals, Algorithms and Applications*. CRC Press, 29–50.

[18] Aurélie Foucquier, Sylvain Robert, Frédéric Suard, Louis Stéphan, and Arnaud Jay. 2013. State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews* 23 (2013), 272–288. https://doi.org/10.1016/j.rser.2013.03.004

[19] Nelson Fumo and M. A. Rafe Biswas. 2015. Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews* 47 (2015), 332–343. https://doi.org/10.1016/j.rser.2015.03.035

[20] E.J. GUMBEL. 1935. Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré* 2, 5 (1935), 115–158. http://www.numdam.org/item/AIHP{_}1935{_}{_}5{_}2{_}115{_}0/

[21] Katherine A. Heller and Zoubin Ghahramani. 2005. Bayesian hierarchical clustering. In *Proc. of the 22nd Int. Conference on Machine learning (ICML'05)*. ACM Press, New York, New York, USA, 297–304. https://doi.org/10.1145/1102351.1102389

[22] Luis Hernandez, Carlos Baladron, Javier M Aguiar, Belen Carro, Antonio J Sanchez-Esguevillas, Jaime Lloret, and Joaquim Massana. 2014. A survey on electric power demand forecasting: Future trends in smart grids, microgrids and smart buildings. *IEEE Communications Surveys and Tutorials* 16, 3 (2014), 1460–1495. https://doi.org/10.1109/SURV.2014.032014.00094

[23] Yaocong Hu, MingQi Lu, and Xiaobo Lu. 2018. Spatial-Temporal Fusion Convolutional Neural Network for Simulated Driving Behavior Recognition. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 1271–1277. https://doi.org/10.1109/ICARCV.2018.8581201

[24] Bertrand Iooss and Paul Lemaître. 2014. A review on global sensitivity analysis methods. *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications* (apr 2014). arXiv:1404.2405 http://arxiv.org/abs/1404.2405

[25] Manar Jaradat, Moath Jarrah, Abdelkader Bousselham, Yaser Jararweh, and Mahmoud Al-Ayyoub. 2015. The internet of energy: Smart sensor networks and big data management for smart grid. *Procedia Computer Science* 56, 1 (2015), 592–597. https://doi.org/10.1016/j.procs.2015.07.250

[26] S W Jefferson. 1998. Modeling Large Forest Fires as Extreme Events Antecedents. 72 (1998).

[27] Shibily Joseph, Jasmin E.A., and Soumya Chandran. 2015. Stream Computing: Opportunities and Challenges in Smart Grid. *Procedia Technology* 21 (2015), 49–53. https://doi.org/10.1016/j.protcy.2015.10.008

[28] H Keemink. [n.d.]. Detecting central heating boiler malfunctions using smart- thermostat data. *TU Delft* ([n. d.]).

[29] Abrar Khameis, Shaikhah Rashed, Ali Abou-Elnour, and Mohammed Tarique. 2015. Zigbee based optimal scheduling system for home appliances in the United Arab Emirates. *Network Protocols and Algorithms* 7, 2 (2015), 60–80. https://doi.org/10.5296/npa.v7i2.7676

[30] Saeed Uz Zaman Khan, Tanvir Hasnain Shovon, Jubayer Shawon, Adeeb Shahriar Zaman, and Saadi Sabyasachi. 2013. Smart box: A TV remote controller based programmable home appliance manager. *2013 International Conference on Informatics, Electronics and Vision, ICIEV 2013* Iii (2013). https://doi.org/10.1109/ICIEV.2013.6572610

[31] Jaime Lloret, Jesus Tomas, Alejandro Canovas, and Lorena Parra. 2016. An Integrated IoT Architecture for Smart Metering. *IEEE Communications Magazine* 54, 12 (dec 2016), 50–57. https://doi.org/10.1109/MCOM.2016.1600647CM

[32] Aurore Lomet, Frédéric Suard, and David Chèze. 2015. Statistical Modeling for Real Domestic Hot Water Consumption Forecasting. *Energy Procedia* 70 (2015), 379–387. https://doi.org/10.1016/j.egypro.2015.02.138

[33] Josip Lorincz, Antonio Capone, and Jinsong Wu. 2019. Greener, Energy-Efficient and Sustainable Networks: State-Of-The-Art and New Trends. https://doi.org/10.3390/s19224864

[34] Zhanyu Ma, Hailong Li, Qie Sun, Chao Wang, Aibin Yan, and Fredrik Starfelt. 2014. Statistical analysis of energy consumption patterns on the heat demand of buildings in district heating systems. *Energy and Buildings* 85 (2014), 664–672. https://doi.org/10.1016/j.enbuild.2014.09.048

[35] Andrea Mauri, Achilleas Psyllidis, and Alessandro Bozzon. 2018. Social Smart Meter: Identifying Energy Consumption Behavior in User-Generated Content. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. ACM Press, New York, New York, USA, 195–198. https://doi.org/10.1145/3184558.3186977

[36] R. Mena, F. Rodríguez, M. Castilla, and M. R. Arahal. 2014. A prediction model based on neural networks for the energy consumption of a bioclimatic building. *Energy and Buildings* 82 (2014), 142–155. https://doi.org/10.1016/j.enbuild.2014.06.052

[37] Krishna Modi and Bhavesh Oza. 2017. Outlier Analysis Approaches in Data Mining. *Ijirt* 3, 7 (2017), 2349–6002.

[38] Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci. 2015. *Introduction to Time Series Analysis and Forecasting*. Wiley–Blackwell.

[39] Office of Electricity Delivery & Energy Reliability. 2016. Advanced Metering Infraestructure and Customer Systems. *Results from the Smart Grid investment grant program* (2016), 98. https://www.energy.gov/sites/prod/files/2016/12/f34/AMISummaryReport{_}09-26-16.pdf

[40] Ripon Patgiri and Arif Ahmed. 2016. Big Data: The V's of the Game Changer Paradigm. In *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 17–24. https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0014

[41] Energie Picardie. 2019. RENOVATION . PICARDIE . FR / PASS- RENOVATION / DYNAMISER-TERRITOIRE-. (2019), 1–6. https://www.pass-renovation.picardie.fr/

[42] Foster Provost and Tom Fawcett. 2014. Authors' Response to Gong's, "Comment on Data Science and its Relationship to Big Data and Data-Driven Decision Making". *Big Data* 2, 1 (2014), 1. https://doi.org/10.1089/big.2014.1516

[43] Jonathan L Ramseur. 2019. U . S . Carbon Dioxide Emissions in the Electricity Sector : Factors , Trends , and Projections SUMMARY U . S . Carbon Dioxide Emissions in the Electricity Sector : Factors , Trends , and Projections. (2019).

[44] B Salina and P Malathi. 2014. An Efficient Data Fusion Architecture for Location Estimation Using FPGA. 3, 1 (2014), 2634–2639.

[45] The Sean developers. 2018. Docker Documentation. 6.1.0.dev0 (2018).

[46] Sandra Sendra, Jaime Lloret, Miguel García, and José F Toledo. 2011. Power saving and energy optimization techniques for wireless sensor networks. *Journal of Communications* 6, 6 (2011), 439–459. https://doi.org/10.4304/jcm.6.6.439-459

[47] Heng Shi, Minghao Xu, and Ran Li. 2018. Deep Learning for Household Load Forecasting-A Novel Pooling Deep RNN. *IEEE Transactions on Smart Grid* 9, 5 (2018), 5271–5280. https://doi.org/10.1109/TSG.2017.2686012

[48] L. Suganthi and Anand A. Samuel. 2012. Energy models for demand forecasting - A review. *Renewable and Sustainable Energy Reviews* 16, 2 (2012), 1223–1240. https://doi.org/10.1016/j.rser.2011.08.014

[49] Utility Analytics Institute. 2017. The Current State of Smart Grid Analytics. (2017), June 13.

[50] G. Vargas-Solar, J. L. Zechinelli-Martini, and J. A. Espinosa-Oviedo. 2017. Big Data Management: What to Keep from the Past to Face Future Challenges? *Data Science and Engineering* 2, 4 (dec 2017), 328–345. https://doi.org/10.1007/s41019-017-0043-3

[51] Kun Wang, Yihui Wang, Yanfei Sun, Song Guo, and Jinsong Wu. 2016. Green industrial Internet of Things architecture: An energy-efficient perspective. *IEEE Communications Magazine* 54, 12 (2016), 48–54.

[52] Xiping Wang and Ming Meng. 2012. A hybrid neural network and ARIMA model for energy consumption forecasting. *Journal of Computers* 7, 5 (2012), 1184–1190. https://doi.org/10.4304/jcp.7.5.1184-1190

[53] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Transactions on Smart Grid* June (2018), 1–24. https://doi.org/10.1109/TSG.2018.2818167 arXiv:arXiv:1802.04117v2

[54] Yang Weng and Ram Rajagopal. 2015. Probabilistic baseline estimation via Gaussian process. *IEEE Power and Energy Society General Meeting* 2015-Septe (2015). https://doi.org/10.1109/PESGM.2015.7285756

[55] Tri Kurniawan Wijaya, Matteo Vasirani, and Karl Aberer. 2014. When bias matters: An economic assessment of demand response baselines for residential customers. *IEEE Transactions on Smart Grid* 5, 4 (2014), 1755–1763. https://doi.org/10.1109/TSG.2014.2309053

[56] Working Group on Big Data Analytics and Machine Learning and Artificial Intelligence in the Smart Grid. 2017. Big Data Analytics in the Smart Grid. *IEEE Smart Grid* (2017).

[57] Jinsong Wu. 2012. Green wireless communications: from concept to reality [industry perspectives]. *IEEE Wireless Communications* 19, 4 (2012), 4–5.

[58] Jinsong Wu, Song Guo, Huawei Huang, William Liu, and Yong Xiang. 2018. Information and communications technologies for sustainable development goals: state-of-the-art, needs and perspectives. *IEEE Communications Surveys & Tutorials* 20, 3 (2018), 2389–2406.

[59] Jinsong Wu, Song Guo, Jie Li, and Deze Zeng. 2016. Big data meet green challenges: Greening big data. *IEEE Systems Journal* 10, 3 (2016), 873–887.

[60] Jingrui Xie, Tao Hong, and Joshua Stroud. 2015. Long-term retail energy forecasting with consideration of residential customer attrition. *IEEE Transactions on Smart Grid* 6, 5 (2015), 2245–2252. https://doi.org/10.1109/TSG.2014.2388078

[61] Junjing Yang, Chao Ning, Chirag Deb, Fan Zhang, David Cheong, Siew Eang Lee, Chandra Sekhar, and Kwok Wai Tham. 2017. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings* 146 (2017), 27–37. https://doi.org/10.1016/j.enbuild.2017.03.071

[62] Yi Wang, Qixin Chen, Chongqing Kang, Mingming Zhang, Ke Wang, and Yun Zhao. 2015. Load profiling and its application to demand response: A review. *Tsinghua Science and Technology* 20, 2 (apr 2015), 117–129. https://doi.org/10.1109/TST.2015.7085625

[63] Wei Yu, Baizhan Li, Yarong Lei, and Meng Liu. 2011. Analysis of a residential building energy consumption demand model. *Energies* 4, 3 (2011), 475–487. https://doi.org/10.3390/en4030475

[64] Hai Xiang Zhao and Frédéric Magoulès. 2012. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews* 16, 6 (2012), 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049

A GLOSSARY

Anomaly detection. Identification of rare observations that raise suspicion after being significantly different from the other observations that can be considered as bad data [53]. Anomaly detection and correction are vital for forecasting since a model with outliers (see definition of outliers below) might result in biased parameters estimation.

Bad data. Refer to missing values or unusual patterns caused by unplanned events during data collection and communication (e.g., abnormal stops or restarts of the smart meter).

Data analytics. Analytical activities that varies along a continuum [47]: (i) descriptive analysis consisting of data visualisation, data mining and aggregation reports targeting the understanding of the data stemming from consumption sensing to decide how to process it; (ii) diagnostic analytics aiming the identification of the cause of given events; (iii) predictive analytics addressing the ability to make probabilistic predictions; (iv) prescriptive analytics that utilises techniques like simulation and decision support to find the optimal strategies that can mitigate future risks.

Data preparation. Phase of the data analytics pipeline that includes data collection and anomaly detection. This phase also deals with outliers that affect the quality of the model used.

Data collection. Refer to harvesting "relevant" and historical values (i.e., not all historical data are useful). Since storage and harvesting changes over time, one has to deal with missing or corrupted data.

Demand response implementation (DR). Change in the regular consumption of electric usage by end-users. This change is due to a response to changes in the price over time or to incentive payments [40].

Internet of Things (IoT). Infrastructure of interrelated devices, mechanical and digital machines, and objects communicating over the Internet. When applied to electric utilities in buildings, it promotes the implementation of a "Smart Building".

Load profiling. Used to determine basic electricity consumption patterns of different costumers' groups by classifying consumers' load curves according to their energy consumption behaviour: (i) direct clustering-based approach with different classification techniques used like K-means[20] [42], hierarchical clustering [61], and self-organising map (SOM) [50]; (ii) indirect clustering includes dimensionality reduction, load characteristics and uncertainty-based methods depending on the features extracted before clustering. Note that most clustering techniques use historical data that requires techniques to deal with the huge amount of streaming data gathered by smart meters.

Outlier. A data point is considered as an outlier when it diverges from an overall pattern on a sample.

Power load analysis. Power analysis performed on the distribution system to ensure balancing and no overloading in any place on the grid. Load analysis results can be further used for load forecast and demand response programs.

Representational State Transfer (REST). Software architectural style that defines a set of constraints to be used for creating Web services. Web services that conform to the REST architectural style, called RESTful services, provide interoperability between computer systems on the internet. RESTful services allow the requesting systems to access and manipulate textual representations of Web resources by using a uniform and predefined set of stateless operations.

Smart meter. Electronic device that records electric energy consumption and communicates data to the electricity supplier for monitoring and billing [22]. The high frequency of data readings opens new possibilities for understanding the electricity demand network [60]. By providing real-time data, a smart meter allows utility providers to optimise energy distribution while allowing consumers to make smarter decisions about their energy consumption and associated carbon impact [10].

Smart metering environment. System that measures, collects, and analyses data collected by meters (i.e., physical variables like gas and electric consumption, temperature, humidity, occupancy, etc.). A smart metering system consists of three main components:

(i) Smart meters installed in households that send data at a specific rate (e.g., each 5, 30 or 60 secs).
(ii) Communication networks to transmit data from and to the smart meters equipped in the households.
(iii) Data management system to store and process data, and send back data like billing information, load forecast, real-time carbon footprint, etc.