Chapter 10
Case Study from the Energy Domain

Dea Pujić(✉), Marko Jelić, Nikola Tomašević, and Marko Batić

Institute Mihajlo Pupin, University of Belgrade, Belgrade, Serbia
dea.pujic@pupin.rs

Abstract. Information systems are most often the main focus when considering applications of Big Data technology. However, the energy domain is more than suitable also given the worldwide coverage of electrification. Additionally, the energy sector has been recognized to be in dire need of modernization, which would include tackling (i.e. processing, storing and interpreting) a vast amount of data. The motivation for including a case study on the applications of big data technologies in the energy domain is clear, and is thus the purpose of this chapter. An application of linked data and post-processing energy data has been covered, whilst a special focus has been put on the analytical services involved, concrete methodologies and their exploitation.

1 Introduction

Big Data technologies are often used in domains where data is generated, stored and processed at rates that cannot be efficiently processed by one computer. One of those domains is definitely that of energy. Here, the processes of energy generation, transmission, distribution and use have to be concurrently monitored and analyzed in order to assure system stability without brownouts or blackouts. The transmission systems (grids) that transport electric energy are in general very large and robust infrastructures that are accompanied by a great deal of monitoring equipment. Novel Internet of Things (IoT) concepts of smart and interconnected homes are also pushing both sensors and actuators into peoples homes. The power supply of any country is considered to be one the most critical systems and as such its stability is of utmost importance. To that effect, a wide variety of systems are deployed for monitoring and control. Some of these tools are presented in this chapter with a few from the perspective of end users (Non-Intrusive Load Monitoring, Energy Conservation Measures and User Benchmarking) and a few from the perspective of the grid (production, demand and price forecasting).

2 Challenges Withing the Big Data Energy Domain

In order to be able to provide advanced smart grid, user-oriented services, which will be discussed further in this chapter, integration with high volume, heterogeneous smart metering data (coming both from the grid side, e.g. placed in power

V. Janev et al. (Eds.): Knowledge Graphs and Big Data Processing, LNCS 12072, pp. 165–180, 2020.
https://doi.org/10.1007/978-3-030-53199-7_10

substations, and from the user side, e.g. installed in homes and buildings) is a prerequisite. To specify, suggest and deliver adequate services to end users (i.e. energy consumers) with respect to their requirements and power grid status, various forms of energy data analytics should be applied by distribution system operators (DSO) and grid operators such as precise short- and long-term energy production and consumption forecasting. In order to deliver such energy analytics, historical energy production data from renewable energy sources (RES) and historical consumption data, based on smart metering at consumer premises and LV/MV power substations, must be taken into account.

The main challenge to providing advanced smart grid services is related to the integration and interoperability of high volume heterogeneous data sources as well as adequate processing of the acquired data. Furthermore, making this data interoperable, based on Linked Data API, and interlinked with other data sources, such as weather data for renewable energy sources (RET) production analysis, number of inhabitants per home units, etc., is essential for providing additional efficient user tailored analytical services such as energy conservation action suggestions, comparison with other consumers of the same type, etc.

Another challenge is related to analysis of grid operations, fault diagnostics and detection. To provide such advanced analytics, real-time integration and big data analysis performed upon the high volume data streams coming from metering devices and power grid elements (e.g. switches, transformers, etc.) is necessary, and could be solved using Linked Data principles. Finally, to support next generation technologies enabling smart grids with an increased share of renewables, it is necessary to provide highly modular and adaptable power grids. In addition, adequate tools for off-line analysis of power system optimal design should be deployed. These analytical tools should also incorporate allocation of optimal reconfiguration of power grid elements to provide reliable and flexible operation as an answer to the changing operational conditions. Tools for planning and reconfiguring power distribution networks consider power station infrastructure and its design, number and capacity of power lines, etc. To provide such advanced grid capabilities, integration with historical power grid data, archives of detected alarms and other relevant operational data (such as data from smart metering, consumption data, etc.) is necessary. Therefore, the main challenge is to provide digested input to the batch-processing, big data analytics for power grid infrastructure planning.

Having all of this in mind, the significance of big data processing techniques is obvious. On the other hand, further in this chapter examples of analytical services will be presented and discussed.

3 Energy Conservation Big Data Analytical Services

Improving quality of life through advanced analytics is common nowadays in various domains. Consequently, within the energy domain, collecting data from numerous smart meters, processing it and drawing conclusions are common concepts in the field of developing energy conversation services. The amount of aforementioned data highly depends on the service's principal use. If the focus is put on just one household, data can be undoubtedly processed using only one computer. Nonetheless, if the scale of a problem is a neighbourhood, municipality or city level, data processing and analytical computations can be taken as a big data problem. Therefore, within this chapter, methodologies for smart energy services are going to be discussed.

3.1 Non-Intrusive Load Monitoring

The first of these is so-called Non-Intrusive Load Monitoring (NILM). NILM was motivated by conclusions, such as those from [70], which claimed that up to 12% of residential energy consumption can be decreased by giving users feedback on how the energy has been used. In other words, by providing the user with information about which of their appliances is using electrical energy and how much, significant savings can be reached. Nonetheless, providing this kind of information would require installation of numerous meters all around households, which is usually unacceptable for the end-user. Therefore, instead of the Intrusive Load Monitoring solution which influences users' convenience, Non-Intrusive Load Monitoring was proposed by Hart in [183] with the main goal of providing users with the same information in a harmless way by aggregating entire household consumption at the appliance level, which can be seen in Fig. 1.

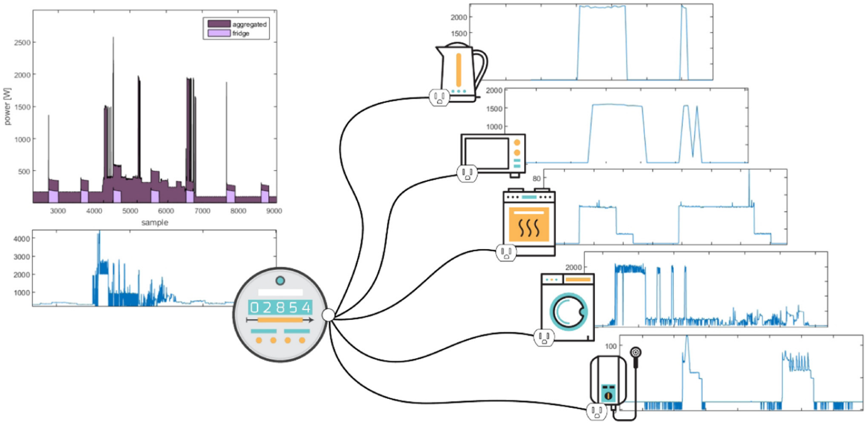

Fig. 1. Non-Intrusive Load Monitoring concept

Having in mind the previous information, two main problems are present within the NILM literature - **classification**, which provides information about the activation on the appliance level, and **regression** for the estimation of the appliance's individual consumption, as shown in the example Fig. 2. As these are some of the most common problems in advanced analytics, typical methodologies employed to address these are leading machine learning approaches, which are going to be presented and discussed further in this section to give an example of the use of applied big data technologies in the energy domain.

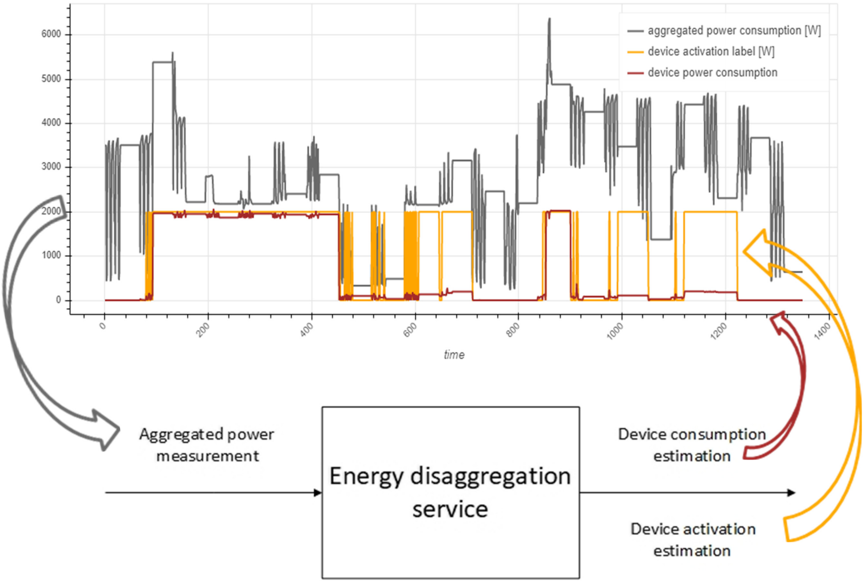

Fig. 2. NILM classification and regression example

As a first step, in this section, the currently present publicly available datasets will be introduced as the basis of data-driven models, which will be discussed further. Depending on the sampling rate, within the NILM literature, data and further corresponding methodologies are usually separated in two groups - **high** and **low** frequency ones. For high frequency, measurements with a sampling time of less than 1 ms are considered. These kind of data are usually unavailable in everyday practice due to the fact that usual residential metering equipment has a sampling period around 1 s and is put as the low frequency group. This difference in sampling rate further influences the choice of the disaggregation methodology and preprocessing approach for the real-time coming data used as the corresponding inputs.

When discussing publicly available data sets, methodologies are not strictly separated in accordance with the chosen sampling rate but rather by the geographical location. In other words, measurements usually correspond to some localized neighbourhood from which both high and low frequency data might be found in the same data set. The first published dataset we refer to is REDD (Reference Energy Disaggregation Data Set, 2011) [256]. It includes both low and high sampling frequency measurements from six homes in the USA. For the first group, both individual and aggregated power measurements were covered for 16 different appliances, allowing the development of various models, which require labeled data. By contrast, high frequency measurements contain only aggregated data from the household, so the developers have to use unsupervised techniques. Another widely spread and used data set published with [238] is UK-DALE (UK Domestic Appliance-Level Electricity) collected in the United Kingdom from five houses. It, again, covers the whole range of sampling rates, and, similarly to REDD, contains labeled data only for those with a sampling period bigger than 1 s. Additional data sets that should be addressed are REFIT [318], ECO (Electricity Consumption and Occupancy) [33], IHEPCDS (Individual household electric power consumption Data Set) [319] for low sampling rate and BLUED [137] and PLAID [145] for the high one[1].

After presenting the available data, potential and common problems with data processing as part of the theme of big data will be discussed. The first one, present in most of the data sets, is the presence of the **missing data**. Depending on the data set and the specific household appliance, the scale of this problem varies. For example, in the case of refrigerators, this is a minor problem which can be neglected because it works circularly, so each approximately 20 min it turns on or off, leading to numerous examples of both active and inactive working periods. By contrast, when, for example, a washing machine is considered, dropping down the sequence of its activation is unacceptable as it is turned on twice a week in a household on average, so it is difficult to collect enough data for training purposes. Therefore, different techniques were adapted in different papers for additional data synthesization from simply adding existing individual measurements of the appliance's consumption on the aggregated power measurements in some intervals when the considered appliance has not been working to more sophisticated approaches such as generative modeling, which was used to enrich data from commercial sector measurements [193].

It is worth mentioning here that characteristics of the data from these different sets significantly deviate in some aspects as a result of differences in location, habits, choice of domestic appliance, number of occupants, the average age of the occupant etc. The NILM literature has attempted to address this **generalization problem**. Even though the problem of achieving as high performance as possible on the testing rather than training domain is a hot topic in many fields of research within Machine Learning (ML) and Big Data, the generalization problem is even more crucial for NILM. As different houses might include different types of the same appliances, the performance on the data coming from the house whose measurements have not been used in the training process might be significantly lower than the estimated one. Additionally, it is obvious that the only application of the NILM models would be in houses which have not been used in the training phase, as they do not have labeled data (otherwise, there would be no need for NILM). Bearing all of this in mind, validating the results from the data coming from the house whose measurements have already been used in the training process is considered inadequate. Thus, it is accepted that for validation and testing purposes one, so called, unseen house is set aside and

[1] http://wiki.nilm.eu/datasets.html.

all further validation and testing is done for that specific house. Nonetheless, the houses covered by some publicly available dataset are by the rule in the same neighbourhood, which leads to the fact that data-driven models learn patterns which are characteristics of the domain rather than the problem. Therefore, separation of the house from the same dataset might be adequate. Finally, the last option would be validating and testing the measurements from the house using a different data set.

State-of-the-art NILM methodologies will be presented later in this section alongside corresponding estimated performance evaluations. Historically, the first ones were Hidden Markov Models and their advancements. They were designed to model the processes with unobservable states, which is indeed the case with the NILM problem. In other words, the goal is to estimate individual consumption in accordance with the observable output (aggregated consumption). This approach and its improvements have been exploited in numerous papers such as [227,245,255,293,294], and [56]. However, in all of the previously listed papers which cover the application of numerous HMM advancements to the NILM problem, the problem of error propagation is present. Namely, as HMM presumes that a current state depends on a previous one, mistakes in estimating previous states have a significant influence on predicting current ones.

Apart from HMMs, there are numerous unsupervised techniques applied for NILM. The main cause of this is the fact that labeled data for the houses in which services are going to be installed are not available, as already discussed. Therefore, many authors choose to use unsupervised learning techniques instead of improving generalization on the supervised ones. Examples of these attempts are shown in [194] where clusterization and histogram analysis has been employed before using the conditional random fields approach, in [344] where adaptation over unlabeled data has been carried out in order to improve performance on the gaining houses, and in [136] where disaggregation was described as a single-channel source separation problem and Non Negative Matrix Factorization and Separation Via Tensor and Matrix Factorization were used. Most of these approaches were compared with the HMM-based one and showed significant improvements. Another approach to gain the best generalization capabilities possible that can be found in the literature is semi-supervised concept in which a combination of supervised and unsupervised learning is present. In [30], self-training has been carried out using internal and external information in order to decrease the necessity of labeled data. Further, [208] proposes the application of transfer learning and blind learning, which exploits data from training and testing houses.

Finally, supervised techniques were widely spread in the literature as well. Currently, various ML algorithms hold a prime position with regards to supervised approaches, as they have proven themselves to be an adequate solution for the discussed problem, as reviewed in [419]. The biggest group currently popular in the literature is neural networks (NNs). Their ability to extract complex features from an input sequence was confirmed to increase their final prediction performance. Namely, two groups stood out to be most frequently used - Recurrent Neural Networks (RNNs) with the accent on Long Short Term Memory (LSTM) [302], and

Convolutional Neural Networks (CNNs) with a specific subcategory of Denoising Autoencoders [239].

After presenting various analytical approaches for solving the NILM problem, it is crucial to finish this subsection with the conclusion that results obtained by this service could be further post-processed and exploited. Namely, disaggregated consumption at the appliance level could be utilized for developing failure detection services in cooperation with other heterogeneous data.

3.2 Energy Conservation Measures (ECM)

When discussing the appeal and benefits of energy savings and energy conservation amongst end users, especially residential ones, it is no surprise that users react most positively and vocally when potential cost savings are mentioned. Of course, when this is the main focus, retrofitting old technologies, improving insulation materials, replacing windows and installing newer and more energy-efficient technologies is usually included in the course of action first recommended. This is mainly because the aspects that are tackled by these modifications are the largest source of potential heat losses and energy conversion inefficiencies. However, there is a significant and still untapped potential for achieving significant energy savings by correcting some aspects of user behaviour.

Besides inefficient materials, bad habits are one of the main causes of high energy loss, especially in heating and cooling applications with the thermal demand being a distinct issue due to the high volume of energy being spent in the residential sector on it. Finding the crucial behavioral patterns that users exhibit when unnecessarily wasting energy is key for efficient mitigation and, therefore, a smart home concept is proposed in order to analyze user behavior and facilitate the necessary changes. In order to obtain data to be able to suggest energy conservation measures, a set of smart sensors should be deployed to monitor various parameters. Some of these sensors could include but are not limited to:

- Smart external meter interfaces (measurement of total energy consumption in real-time);
- Smart electricity plugs and cables (measurement of energy consumption per appliance in real time and possibility of on/off control);
- Smart thermostats (measurement and continuous control of reference temperature and possibly consumed energy);
- Occupancy sensors (measurement of occupancy and motion and ambient temperature also);
- Window sensors (measurements of open/close status of windows and doors and ambient temperature also);
- Volatile organic compound (VOC) sensors (measurement of air quality and ambient temperature)

In some cases where installing smart plugs and cables is not deemed to be economical, a NILM algorithm described in Subsect. 3.1 can be employed in order to infer individual appliance activity statuses using only the data from the external meter. When widespread deployment of such sensors is being done, the amount of data that should be collected, stored and processed quickly grows due to the fact that multiple sensors are to be deployed in each room and that each of the sensors usually reports multiple measurements (e.g. the window sensor reports the temperature besides the open/close status, but also has a set of utility measurements such is the network status strength, battery status, etc. which should also be monitored as they provide crucial data regarding the health of the device itself). Therefore, efficient solutions, possibly from the realm of big data, should be employed in order to facilitate efficient storage and processing of data as the problematic user behavior is time-limited and should be pointed out to the user in due course while a problematic event is ongoing.

A small-scale use case of such a system was tested on around two dozen apartments in the suburbs of Leers, France with the proposed architecture of the system illustrated in Fig. 3. Using such an architecture, the back-end of the system that employs a MySQL database for static data storage regarding the apartment IDs and custom notification settings in conjunction with an ontology for storing room layouts and detailed sensor deployment data provides support for the main ECM engine that analyses data from the real-time IoT-optimized NoSQL Influx database and sends push notifications to the end users notifying them of energy-inefficient behaviour by cross-correlating different measurements from different sensors. For example, when a heating or cooling device is observed to be turned on in an unoccupied space, the user is warned. If the user acts upon such information and resolves the issue, the notification is dismissed automatically, or if the user does not react and the problematic event goes unresolved, he or she is re-notified after a predefined period of time. These events are analyzed

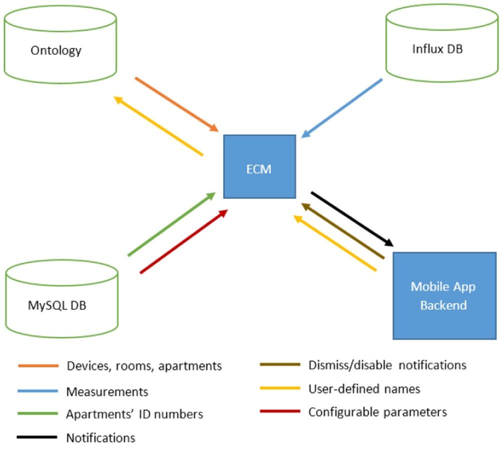

Fig. 3. Proposed architecture of a small-scalle ECM system

with different scopes for individual rooms but also for entire apartments. Also, since smart sensors are already deployed, the energy conservation analysis can also be extended to regard security (no occupancy whilst a door or window is open) and health (poor air quality and windows closed) aspects also. Of course, each event is analyzed separately and appropriate notifications with corrective actions are issued to the end user.

3.3 User Benchmark

Besides the most obvious motivating factor of energy savings – monetary savings – another factor that can greatly impact users' behavior is social pressure. Namely, in a hypothetical scenario where different users were placed in a competition-like environment where the main goal is to be as energy-efficient as possible or, in other words, where each user's score is determined by how efficiently they consume energy, those users would be more likely to strive to perform better and hence consume energy in a more environmentally friendly way. In order to facilitate such an environment, a benchmarking engine has to be developed in order to provide an algorithm that would rank the users.

[81,113] and [329] in the literature point out that the benchmarking procedures in the residential sector have long been neglected in favor of industrial applications. Different algorithms and technologies proposed as core include:

- Simple normalization
- Ordinary least squares (OLS)
- Stochastic frontier analysis (SFA)
- Data envelopment analysis (DEA)
- Simulation (model-based) rankings
- Artificial neural networsk (ANNs)
- Fuzzy reasoning

with related literature [171] offering several dozens of additional related algorithms for multi-criteria decision making (MCDM). The applications of the aforementioned algorithms found in the literature are generally focused on schools, other public buildings and offices, with very few papers, such as [259,291] and [461], analyzing the residential sector.

One of the most prominent standards in energy efficiency ranking is the acclaimed Energy Star program [182], which rates buildings on a scale from 1 to 100 based on models and normalization methods of statistical analysis performed over a database from the US Energy Information Administration (EIA). However, the Energy Star rating does not take into account dynamic data obtained by observing the ongoing behavior of residents. This is where the concept of an IoT-powered smart home can provide a new dimension to energy efficiency benchmarking through real-time analysis of incoming data on how people use the space and appliances at their disposal.

The basis of every ranking algorithm is a set of static parameters that roughly determines the thermal demand of the considered property. These parameters generally include: total heated area, total heated volume, outward wall area, wall thickness, wall conductivity or material, number of reported tenants. This data generally is not massive in volume and is sufficient for some elementary ranking methods. However, an energy efficiency rating that only takes into consideration this data would only have to be calculated once the building is constructed or if some major renovations or retrofits are being made. As such, it would not be able to facilitate a dynamic competition-based environment in which users would compete on a daily or weekly basis on who is consuming their energy in the most economical way.

Given the reasoning above, the static construction and occupancy parameters are extended with a set of dynamic parameters that are inferred based on sensor data collected by the smart home. This data could, for example, include: total consumed energy, occupancy for the entire household, cooling and heating degree days, responsiveness to user-tailored behavior-correcting messages, alignment of load with production from renewable sources, etc. As these parameters are changing on a day-to-day basis, their dynamic nature would provide a fast-paced source that would power the fluctuations in energy efficiency scores of individual users and ultimately help users to see that their change in behaviour has made an impact on their ranking. Also, it is worth mentioning that when users within a same micro-climate are to be ranked, using heating and cooling degree days may prove to be redundant as all users would have the same parameters in this regard. Therefore, this data can be augmented using indoor ambient temperature measurements in order to monitor overheating in winter and overcooling in summer.

The most important procedure that should be conducted within user benchmarking solutions in order to provide a fair comparison between different users with different habits and daily routines is to provide a so-called normalization of consumed energy. This means that, for example, larger consumers should not be discriminated just based on higher consumption; rather, other factors such as the amount of space that requires air conditioning or the number of people using the considered space should be taken into account. In this regard, simply dividing the total consumed energy by the, for example, heated area provides a good first estimate of how energy-efficient different users are per unit of surface, but also implies that a linear relation between area and energy is assumed, which might not be their inherent relationship. In order to mitigate against this issue, vast amounts of data should be collected from individual households using IoT sensors and analyzed in order to either deduce appropriate relations required for normalization or to provide a basis for the aforementioned algorithms (DEA, SFA, etc.), which assign different weights to each of the parameters taken into account.

4 Forecasters

Following the widespread deployment of renewable sources such as wind turbines, photovolotaic panels, geothermal sources, biomass plants, solar thermal collectors and others, mainly as a result of various government-enforced schemes, programs and applicable feed-in tariffs, the stability of the grid has been significantly compromised. The integration of these novel sources has proven to be a relatively cumbersome task due to their stochastic nature and variable production profile, which will be covered in greater depth in Subsect. 4.2. Since the production of most of these sources is highly correlated with meteorological data (wind turbine production with wind speed and photovoltaic production with irradiance and cloud coverage), legacy electrical generation capacities (coal, nuclear and hydro power plants) which have a significantly shorter transient between different states of power output have to balance the fast-paced variations in generation that are a byproduct of the introduction of renewable sources. Since total generation is planned in order to be able to fulfill the total demand that will be requested, being able to know beforehand how much energy will be required in the future and how much energy will be available can provide a basis for potential energy and cost savings through optimal resource planning.

4.1 Demand Forecaster

Given the importance of demand forecasting, it is expected that this topic will be covered by more than a few authors in their published research. However, even though there is a noticeable number of publications in this regard, the topic of energy demand forecasting and the methods used for its estimation still appear to be under-explored without a unified proposed approach and most of the studies being case-specific. In that regard, a probabilistic approach for peak demand production is analyzed in [322], an autoregressive model for intra-hour and hourly demand in [450] and ANN-powered short-term forecasting in [401]. Short-term forecasting is also analyzed whilst making use of MARS, SVR and ARIMA models in [9] and [463] presenting a predictive ML approach. Deep learning frameworks are discussed by [34] and [466]. DSM in connection with time-of-use tariffs is analyzed by [200] and simultaneous predictions of electricity price and demand in smart grids in [314].

Some authors like [105,149,195] and [12] also discuss demand forecasting but place the focus of their research on the predictors that can be used to predict and correlate with the demand values. In this regard, [486] analyzes the correlation of indoor thermal performance and energy consumption. However, again, very few studies focus on residential users, i.e. households and apartments, especially with regard to dynamic data that depicts the ongoing use of that household.

In line with what other authors have noted in their work, the crucial factors that affect demand and that are to be taken into account when building predictive models are the meteorological conditions of the analyzed site. In essence, this correlation is not direct, but rather the temperature, wind speed and direction and irradiance have a significant impact on the use of heating and cooling devices, which are usually the largest consumers of energy in residential households without district heating and cooling. Besides, the current season of the year in moderate climates greatly determines what climatic conditions can be expected, and, therefore, the geographic properties of the analyzed site have to be taken into account since it is the location that determines how severe the seasonal variations in climatic conditions will be. As for the static data, the total floor space or heated volume are also said to be closely correlated with total consumption, but cannot be used to dynamically estimate demand with high time resolution. Here is where large volumes of IoT sensor data collected directly from homes can be of great help in increasing the precision of predictive models. Namely, indoor ambient temperature coupled with outdoor meteorological conditions with live occupancy data in real time can provide a precise short-term estimation of the consumption profile. Furthermore, if past behaviour is taken into account (in the form of previous demand curves both as an average over a larger time period in the past and the more current ones from the previous couple of days) with current day indicators (i.e. whether it is a working day or weekend/holiday), relatively precise hourly and possibly even inter-hourly profiles can be generated.

The presence of smart measuring devices in the form of smart plugs and cables which report real-time consumption per appliance in a home, or their substitution with an NILM algorithm as described in Subsect. 3.1 where bad performance due to insufficient generalization is not an issue, provides the possibility of predicting demand on a per-appliance level. This approach is scarcely depicted in contemporary research articles with only a few papers like [28,312] and [226] exploring this subject. Alternatively, the problem of demand forecasting is most often approached from an aggregated perspective, through the prediction of neighbourhood, city or state-level consumption, with data availability generally being the driving factor that ultimately decides what type of demand will be estimated. Time series from Figs. 4, 5 and 6 illustrate the different dynamics of the demand signals from a single appliance, all appliances of one home and several aggregated homes. Since each of these applications usually requires different levels of prediction precision, the raw data used for these illustrations was averaged with different sample intervals (15 s, 60 s and 15 min) in accordance with the appropriate use case.

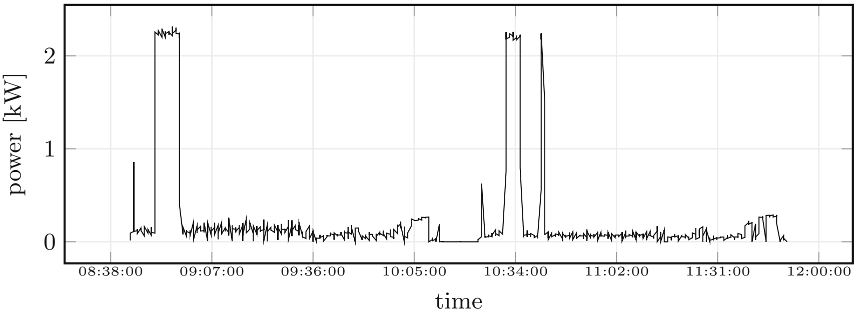

Fig. 4. Typical washing machine demand profile with 15 s averages (showing what appear to be two activations in the span of 4 h)

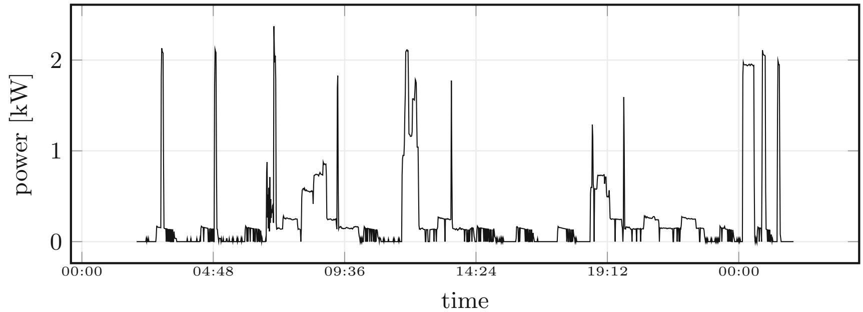

Fig. 5. Total household demand profile with 60 s averages (showing several appliance activations during a full 24-h period)

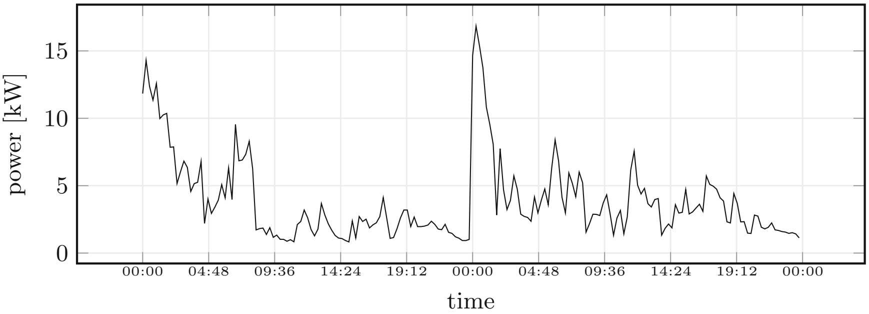

Fig. 6. Aggregate household demand profile with 15 min averages (showing consumption for 2 days with time-of-use tariff)

4.2 Production Forecaster

It has already been mentioned that energy planning is crucial for grid stability, and that it highly depends on the forecast renewable energy sources (RES) production. Therefore, in this subsection different methodologies used for production forecasting are going to be covered as well as their relation to the field of big data.

The production of RES technologies is highly influenced by weather conditions. For example, there is very high dependency between PV production and solar radiation, similar to the relationship between wind turbines and wind speed and direction. In Table 1, the selection of weather services is given followed by their main characteristics. Namely, depending on the practical application, production forecasters can have different time resolutions and horizons, which dictates the necessary weather forecast parameters. Therefore, from the abovementioned table, it can be seen that Darksky can provide estimations in terms of minutes, whilst its horizon, as some kind of compromise, is only 7 days. Additionally, depending on the approach, historical weather data might be necessary for the purpose of the training process, as, currently, the most popular approaches in the field of RES production are data-driven algorithms. Finally, the choice of weather service highly influences its price. All of those characteristics can be found in the table.

Table 1. Overview of forecasting data providers

Name	Min. forecast resolution	Max. horizon [days]	Historical data	Free up to	Coverage
OpenWeatherMap	hourly	30	Yes	60 calls/minute	Global
Weatherbit	hourly	16	Yes	500 calls/day	Global
AccuWeather	hourly	15	prev. 24 h	50 calls/day	Global
Darksky	minute	7	Yes	1000 calls/day	Global
weathersteak	hourly	14	Yes	1000 calls/month	
Yahoo! Weather	hourly	10	No	2000 calls/day	Global
The Weather Channel	15 min	30	Yes		Global
World Weather Online	hourly	15	Yes	Not free	Global

Depending on the practical application apart from input weather parameters developed methodology varies, as well. For the use cases in which few measurements are available, physical models are usually chosen. These models are based on mathematical models and are usually deployed when there are not enough real world measurements. These models are characterized with the lowest performances in comparison with the following ones, but exist in cases of missing data. This methodology is present in the literature for various RES such as photo-voltaic panels (PVs) [115,334], wind turbines (WTs) [273] and solar-thermal collectors (STCs) [80,394]. However, even though they do not require huge amounts of measurements, physical characteristics such as number of solar panels, position of panels and wind turbines, capacity etc. are needed and sometimes, again, inaccessible. Taking into account suppliers' tendency to equip the grid with numerous IoT sensors nowadays, the necessity of physical models is decreasing, leaving room for data-driven models, which are a more important part of this chapter and within the field of big data.

Currently the most popular and explored topic in the field of RES production forecasters is statistical and machine learning (ML) based techniques, which were proven to achieve higher performances but require substantial amounts of data. Nonetheless, bearing in mind that a huge amount of big data is currently available in the energy domain, these approaches are not common only amongst researchers but also in real practice. The first group that stands out are the statistical autoregressive methodologies SARIMA, NARIMA, ARMA, etc. [437]. They are followed by probabilistic approaches, such as in [452]. Finally, neural networks and machine learning-based approaches are proven as one of the most suitable choices [205,236,453], similar to numerous other fields.

Apart from the similar inputs regarding weather parameters and applied models for RES production forecasters, all of the methodologies are dependent on the estimation time horizon. Depending on the practical application, the orders of magnitude can range from minutes to years. Further post-processing of the obtained forecast results is another important factor. Apart from the grid control and stability, from the perspective of big data the analytical tool developed on top of the results provided by the forecaster could be exploited for failure and irregularity detection in the system together with its high level metadata. By contrast, outputs with the big time horizon could be seen as adequate for extracting conclusions on a yearly basis using big data tools already presented in this book.

4.3 Pricing Prediction

Another important application of prediction algorithms in the energy domain are price predictions. As energy sectors worldwide are becoming increasingly deregulated, variable pricing in energy trading is becoming increasingly prominent with some envisioning a not-so-distant future where the cost of energy in the wholesale and maybe even retail markets will be changing every 15 min while the standard nowadays is usually hourly changes at most. Having accurate predictions of wholesale market prices presents key information for large-scale energy traders because it provides an insight into future trends in the same way as stock price predictions do and allows for sound investment planning.

Wholesale price variations greatly impact retail prices, which, in turn, have a key influence on the shape of the expected demand curve from end users. Moving from fixed pricing to first time-of-use tariffs and later hourly variable pricing has allowed for energy retailers to have granular control of load levels through what is essentially implicit demand response (DR) where load increase or decrease events are defined by the current prices. Energy prices are also influenced by the availability of renewable sources. For example, systems with high PV penetration tend to have lower prices during mid-day production peaks to try and motivate users to consume more energy when there is a surplus in the system. In that way, demand predictions, production predictions and pricing productions are mutually interconnected in such a way that should result in a balanced system of equal supply and demand.

5 Conclusion

The brief overview laid out in this chapter provides an insight into some potential applications of big data-oriented tools and analytical technologies in the energy domain. With the importance of climate change mitigation growing by the day, the number of solutions working towards increasing energy efficiency and responsible energy use is only expected to rise. As such, this domain provides an interesting and challenging realm for novel research approaches.