Data Analytics Based Demand Profiling and Advanced Demand Side Management for Flexible Operation of Sustainable Power Networks

A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2019

Miss Jelena Ponoćko, B.S., MSc

School of Electrical and Electronic Engineering

Contents

Word count: 58,431

List of Figures

List of Tables

List of Abbreviations

AMR – Automatic Meter Reading

ANN – Artificial Neural Network

ARIMA – Autoregressive Integrated Moving Average

AWFE – Absolute Weighting Factor Error

CDF – Cumulative Distribution Function

CML – Customer Minutes Lost

CO_2 – Carbon-dioxide

CVR – Conservation Voltage Reduction

DD – Demand Decomposition

DDLC – Decomposed Daily Load Curve

DER – Distributed Energy Resources

DF – Demand Forecasting

DG – Distributed Generation

DLC – Daily Load Curve

DNO – Distribution Network Operator

DR – Demand Response

DRFM – Demand Response Flexibility Market

DSM – Demand Side Management

DSO – Distribution System Operator

ENS – Energy Not Supplied

EV – Electric Vehicle

GPRS – General Packet Radio Service

GSP – Grid Supply Point

GUI – Graphical User Interface

HVAC – Heating, Ventilation, and Air Conditioning

HP – Heat Pump

IM – Induction Motor

KPI – Key Performance Indicator

LED – Light-emitting diode

LI – Linear Interpolation

MAPE – Mean Absolute Percentage Error

MLP – Multilayer Perceptron

NILM – Non-Intrusive Load Monitoring

OHL – Overhead line

OLTC – On-load Tap Changer

OPF – Optimal Power Flow

PAR – Peak to Average Ratio

PDF – Probability Density Function

PF – Power Factor

PMU – Phasor Measurement Unit

PSO – Particle Swarm Optimisation

QoS – Quality of Service

RAE – Relative Absolute Error

RES – Renewable Energy Sources

RMSE – Root Mean Square Error

RRSE – Root Relative Squared Error

SCADA – Supervisory Control and Data Acquisition

SM – Smart Meter

SoC – State of Charge

SMPS – Switch-mode Power Supply

SOM – Self-organising Maps

SQL – Structured Query Language

TCL – Thermostatically Controlled Load

TLP – Typical Load Profile

TSO – Transmission System Operator

ToU – Time of Use

VAR – Volt Ampere Reactive

VG – Virtual Generator

VoLL – Value of Lost Load

VPP – Virtual Power Plant

VSI – Voltage Stability Index

WAkNN – Weight Adjusted k-Nearest Neighbour

WF – Weighting Factor

WFE – Weighting Factor Error

ZIP – Constant Impedance Z, constant current I, constant power P

Abstract

Title: Data Analytics Based Demand Profiling and Advanced Demand Side Management for Flexible Operation of Sustainable Power Networks

Miss Jelena Ponoćko, The University of Manchester, April 2019

With the evolution of smart grid paradigm and the consideration of demand side management (DSM) as one of the flexibility providers in networks with renewable generation, the accurate assessment or prediction of demand profile and advanced DSM are becoming essential. This thesis is contributing to both, an accurate demand profiling and advanced use of DSM to facilitate flexible and secure network operation. It starts by discussing the data and information needed in the future distribution network to facilitate flexible network operation. It then illustrates the benefits of using advanced data mining techniques (artificial neural networks) for better observability of demand in the distribution network with a limited number of smart meters. The first part of the thesis thus illustrates how the flexibility and composition of aggregated demand can be assessed/forecast with very limited information coming from the end users. Once the composition of demand is available, one can assess with high confidence what portion of demand is flexible, what types of load that portion includes (e.g., cold appliances, heaters, etc.), and when and where (at which buses in the network) it should be shifted/curtailed. This enables "tailoring" the DSM program and incentive system to the available size and type of flexible loads in an area. At the same time, it allows a more confident prediction of the outcome of the DSM program (the resulting load curve). Furthermore, it facilitates indirectly a more accurate modelling of demand over a period of time. The second part of the thesis focuses on the use of the information about demand composition, which is first used to model load at each network bus as a composite load model, and then to study different effects of DSM on network operation. Wide-scale DSM involving numerous flexible load buses in the network changes not only the total demand in that area at given time, but also its composition at individual buses, i.e., the shares of different components of the composite load model. This change in demand profile could influence both, the steady state network operation (critical network loading, losses, etc.) and its dynamic performance (voltage and angular stability of the system following a disturbance). Therefore, the second part of the thesis demonstrates how DSM program can be optimally planned hours, or day ahead, across the network, taking into account forecast demand composition and demand flexibility at each bus, in order to meet the requirements of the network operator (e.g., facilitating efficient use of available renewable resources), and at the same time maintain the relevant steady state and/or dynamic performance indicators of the network at the level they were before deployment of the DSM program.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgment

I would like to thank my supervisor, Professor Jovica V. Milanović, for the wonderful opportunity to work in his research group, in room C26 in Ferranti Building. His trust, patience, support and wholehearted dedication to work have been a great source of inspiration and motivation to work hard and always look up ("Celui qui regarde vers le haut ne peut jamais avoir le vertige". – M. Kundera).

My special thanks are extended to my colleagues from the room C26, the current ones and those who left in the meantime to pursue academic or industrial career, for making me feel welcome from the very first day, for their kindness and generous help and advice whenever I needed it. I would also like to thank my friends and colleagues in Ferranti Building, for their friendship and interesting discussions on all possible life and work-related topics. Without the people I met during my PhD, this journey would not have been nearly as enjoyable.

A special acknowledgment should be given to the two EU Horizon 2020 projects, namely NOBEL GRID project and CROSSBOW project, which gave me the opportunity to work on large collaborative projects between academia and industry, and gain irreplaceable skills and experience. Thanks to the work on these projects and support received from The University of Manchester, I was able to fund my PhD research.

Finally, I can never thank enough my greatest supporters, Jagoda, Čeda, Bogdan and Adrien, for always being there, even when we are hundreds and thousands of kilometres apart. They give a meaning to every achievement of mine and I dedicate this thesis to them.

In Manchester, April 1st 2019

To my family and Adrien

1 Introduction

1.1 Introduction

This chapter introduces the two main areas of the research presented in this thesis, with some theoretical aspects. It provides an overview of the past work in these areas, with respect to both research and real-world practices. After identifying some of the gaps in previous research on the topic, the main aims and objectives are defined, followed by extracting the main contributions of this thesis.

1.2 Background

The two evolutionary changes in power industry paving the way to the smart grid concept are the move towards low carbon operation of the power network and the introduction of information and communication technologies (ICT). These changes have been happening at all stages and all aspects of the electricity generation, transmission and distribution. Due to the intermittency of renewable generation, whose share is constantly growing at all voltage levels, the need for higher flexibility of the network operation has been raised, and thus the technologies such as energy storage and demand response (DR) came into the focus. Power system flexibility can be defined in different ways, one of them being “the ability to adapt to dynamic and changing conditions, for example, balancing supply and demand by the hour or minute, or deploying new generation and transmission resources over a period of years” [1]. Activation of the demand side, i.e., more active participation of demand in system operation and control, is one of the main features of the smart grid. Instead of the traditional load-following generation approach, demand is becoming more flexible and adjustable to the available generation, which, due to the volatility of its renewable part, is becoming less controllable than before. Furthermore, with the increasing presence of and reliance on the ICT in power network and monitoring systems in general for enhanced network observability, processing of large amount of diverse data streams and extraction of relevant knowledge from the data requires utilisation of data mining techniques. Data mining has two benefits; in case of big data it enables fast extraction of useful knowledge, while in the case of insufficient observability, it derives knowledge from limited available data, i.e., it enables learning from the past or from similar cases. Addressing the issues of network flexibility by means of previously not harnessed services and technologies and efficient handling of data and knowledge extraction became therefore top priorities for smart grid development in general.

Although the smart grid concept is being developed at all levels of the power system, this thesis focuses on distribution network (DN) and examines the extent to which activation of the demand side flexibility can facilitate daily operation of the DN, while contributing to an extent to the transmission network operation as well. The two prerequisites for utilising demand flexibility are existence of controllable demand and ICT infrastructure for improved demand observability. Since the power system is operated in close to real-time, the traditional load monitoring systems (electricity meters), with readings performed several times a year, have become obsolete for the demand side management (DSM) requirements. It should be noted that DSM and DR will be used interchangeably in the text, even though DSM is a more general term, while DR commonly refers to price-based programs. Increasing number of installed smart meters (SMs) in residential districts around the world will enable better observability of the end-users’ behaviour and their potential to participate in the network daily operation. Higher granularity of low-level consumption data in the future distribution grid will bring benefits to both consumers and the distribution system operator (DSO). On the one hand, smart metering will facilitate awareness of consumers about their daily consumption and enable them to make savings by reacting to price signals or various types of incentives triggered by their electricity supplier. On the other hand, SM data will provide information to the DSO about individual load profiles, enabling more advanced profiling of consumers in different areas and at different levels of aggregation. Authors in [2] pointed out that the analysis of SM data can be applied for load pattern recognition, assessing DR potential, tariff design, load forecasting or sociodemographic identification of end-users.

There is a wide area of research focusing on the application of data mining methods to extract useful knowledge from data coming from readily available and new types of monitors, thus enhancing the observability of the demand side and the distribution network in general. It has been estimated that monitoring of all the LV networks in the UK would cost £2 billion [3]. Therefore, data mining should be used to obtain relevant information from small populations of data coming from new meters and maximise the use of already existing, to an extent limited number of monitors in the network. Following this line of thought, around $3.3 million was invested in China for big data analytics-related projects in electrical engineering, covering, among others, energy forecasting, equipment monitoring and renewable integration [2].

The effectiveness of DSM actions largely depends on the flexibility of the demand side. Until now, mostly large industrial users have been included in DR programs [4]. According to [5], there is more than 500 MW of DR capacity for short-term operating reserve in the UK. Nevertheless, there is a significant, though mainly untapped potential for DR in residential area. In the US, it is estimated that the participation of residential customers in DR might bring up to half of the total peak reduction [6]. Changing the electricity consumption habits of residential users would have significant environmental implications, e.g., [7] reported that residential demand contributes to a quarter of global greenhouse gas emissions. Taking the UK as an example, residential (domestic) sector is the largest final user of electrical energy, presenting around 30% of overall consumption, followed by industrial and commercial sector accounting for 26% and 21% of the total consumption, respectively, see Figure 1.1 [8]. As the impact of individual smaller consumers is negligible, DR potential (i.e., load flexibility) of an aggregated group of users should be investigated instead. The advantages of the "aggregate and dispatch" DSM model are that it overcomes uncertainties of individual units and requires minimum local observability [9].

Aggregators will be very valuable and potentially influential actors in the future distribution power system as they will represent groups of numerous small end-users as "bulks" of users with increased influence. In other words, changing the load pattern of a larger number (e.g., hundreds or thousands) of end-users simultaneously will have much larger effect on the network than changing electrical patterns of individual users.

Another benefit of having aggregators is provision of anonymity of the end-users and their daily and seasonal behaviour through aggregation. The aggregator could collect relevant operational information from dispatchable loads and serve as a mediator between a utility/network operator and individual loads, as discussed in [10]. Possible candidates for dispatch at domestic, individual customer level, include: dishwashers, washing and drying machines, electric water heaters, heat pumps, heating, ventilation and air conditioning (HVAC) with thermal storage, battery chargers and electric vehicles (EVs). Some loads can be deployed for demand dispatch with a relatively quick response and without significant effect on end-users' commodity, while the other can be responsive to changes in grid frequency, in which case they respond automatically, with no intervention by the aggregator.

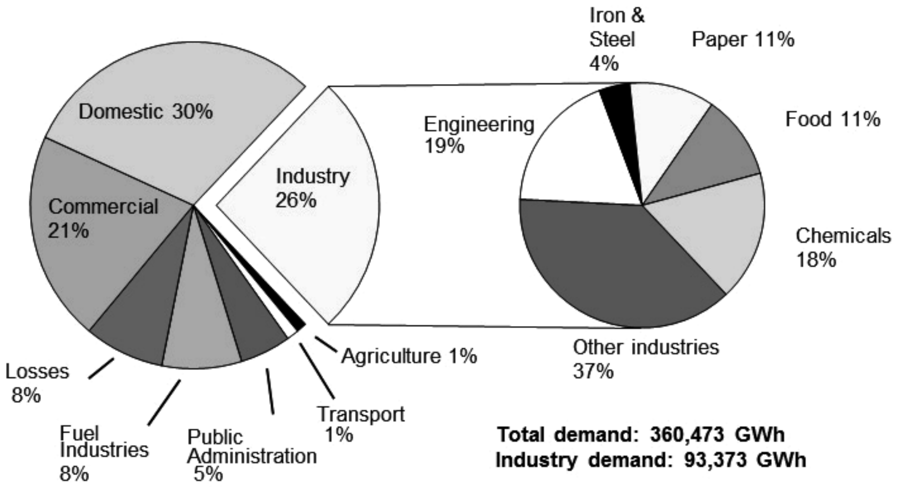


Figure 1.1 Electricity demand by sector in the UK 2014 (adopted from [8])

1.3 Motivation

In the majority of cases, DSM is triggered by the DSOs, suppliers and aggregators [11]. DSOs commonly comprise residential and small and medium size commercial users, while aggregators collect flexibility from these or large commercial and industrial users. As the effect of aggregated DR at the distribution/aggregation level can be significant and transmitted to the transmission level, the network performance indicators (such as network losses or different aspects of frequency and voltage stability) should be recorded and maintained, if not improved during and after control actions. Any planning of DSM program should therefore rely on several inputs:

- The aim of the DSM: load shaping, minimisation of cost or carbon-dioxide (CO_2) emission, etc.;
- Available demand side flexibility: the size and type of controllable loads within the total demand, i.e., the composition of demand and its change during the time;
- Preservation of network performance in cases of wide-scale DSM: minimisation of losses, maintenance of voltage levels or loadability of the network, etc.

The available demand flexibility can be assessed or forecast if appropriate monitoring systems are in place. The two main pieces of information that define load flexibility in every time step are the size of demand (in MW, for example), and the size (share) of controllable (flexible) loads within the total demand. Observability of the end-users is enhanced by SMs, which at the moment have limitations with respect to reporting demand-side flexibility. In other words, there is no widely-available information about the amount of controllable demand at different times of the day and different seasons of the year. Nevertheless, existing SMs can be highly useful for classifying daily load patterns and tariffing purposes. Due to the lack of detailed demand observability, there is a need for alternative approaches that could result in satisfactory accuracy in assessing demand flexibility with limited monitoring data. This thesis therefore suggests a data mining-based approach for advanced demand profiling, i.e., forecasting demand composition in the presence of limited load observability.

Network performance, as the third prerequisite for DSM planning, is estimated based on different indicators (voltage, frequency, line loading, etc.). In this thesis, network performance is observed through steady-state voltage stability in distribution network. Voltage instability (also referred to as load instability) in the distribution system may spread to the transmission system and cause a major blackout [12]. During a heavily loaded condition, even a relatively small but sudden increase in demand can result in voltage instability [13]. Load margin, as a voltage stability indicator, will be of a particular interest once the proliferation of large residential loads becomes significant. The penetration of EVs, for example, could double the current distribution network load, especially at peak load hours [14]. The UK transmission system operator, National Grid, envisages up to 11 million EVs by 2030 and 36 million by 2040 [15]. In addition, up to 60% of homes in UK should be using heat pumps by 2050.

In a large-scale DSM scenario, load flexibility may be harnessed at different buses of the transmission network, by either aggregators or DSOs at the grid supply point (GSP). In this scenario, time shift of large portions of flexible loads across network buses (and consequently space-shift of demand), will change not only the total demand in different regions of the network, i.e., the power flow, but also demand composition (shares of static and dynamic loads) at network buses. Demand size and composition have been repeatedly shown to have crucial impact on the nature of dynamic response of demand following a network disturbance [16, 17], and consequently the impact on the overall network voltage and angular stability. Therefore, the study presented in this thesis considers two network performance-based limitations of DSM, namely distribution network load margin and the composition of aggregated distribution network demand (the demand "seen" from the transmission network).

1.4 The beneficiaries of the research

The results of demand profiling are of critical importance to the DR responsible entity, whether it is a DSO, electricity supplier or an aggregator. Obtaining as much and as accurate as possible information about the load, or in other words obtaining a load profile, is crucial for the studies of direct load control, DR programs, design of tariffs and involvement of local generation [18]. An important part of load profiling is flexibility profiling, i.e., assessment of the size of controllable (shiftable/deferrable) load within the total load. This information can reduce the uncertainty of the actual (available) flexibility of the demand side as a response to the signal sent during a DR program. Ability to forecast aggregated demand and the size of controllable load facilitates assessment of the actual capacity for operating reserve and energy services coming from the demand side (these will be detailed in Section 1.5.3). Flexibility profile of aggregated customers is more predictive than the profile of individual customers, which is highly random. Furthermore, load can be disaggregated (decomposed) into load categories, such as resistive loads, induction motors, lighting, etc., in order to obtain a more detailed insight into the types of load utilised on a daily or seasonal basis.

The profiling can be performed in two dimensions:

- Time: observing the change in the size of controllable load within the total load over a day or season;
- Space: observing the size of controllable load over a distribution network. In this case, different distribution network buses (e.g., 33/11 kV or 33/6.6 kV in the UK) will have different flexibility potential, depending on their load mix (namely residential, industrial or commercial users).

Information about the size and composition of controllable demand allows for efficient and confident DR planning, and enables participation of aggregated end-users in network daily operation by forecasting demand flexibility as one of the distributed energy resources (DER). This further drives definition of appropriate incentives that can be "tailored" according to the demand profile of the group of users. For example, incentive-based system can be introduced in areas of the network where there is a high share of wet appliances (e.g., washing machines). Following the direct load control scenario, where certain load categories are equipped with smart controllers (e.g., electric water heaters), load disaggregation would provide information on the amount and profile of the disposable controllable load. Programs such as conservation voltage reduction (CVR), which are considered as non-intrusive DR, can be introduced in networks with a high share of loads that can be modelled as constant impedance (these loads have the highest sensitivity of consumption to voltage changes). In CVR programs, load reduction of 0.5% per 1% of voltage reduction were observed during winter peak times, as reported in [19].

Once the observability of demand is high enough to allow for more confident short-term planning of DR program and its outcome, analysis of network performance indicators should also be taken into account to make sure the operation of the network is not endangered due to the changes in demand and its composition. Although not very common in the literature, voltage stability of distribution network and microgrids has been analysed in the past [12, 20-22]. It has been reported that the key factors affecting voltage stability of these networks are DER limits and sensitivity of the loads to voltage variation [23]. Therefore, DSM action resulting in changing composition of loads and such potentially the nature of their dependence on voltage could affect voltage stability of the network. The loading limit of the network as a network performance indicator related to voltage stability is thus suitable to illustrate the possible effects of aggregated DR.

Finally, multi-objective DSM, relying on the aforementioned prerequisites (load manipulation, available demand flexibility and network performance) brings benefits to both the DSM provider and the entity requesting it, for example the TSO. It enables confident planning of the DR outcome (changes in loading curve) on one side, and preservation of network performance, in this case loadability of the distribution network, on the other.

1.5 Review of the past work in the area

1.5.1 Demand observability

As previously mentioned, one of the requirements for successful DSM programs is the enhanced observability of the demand side. Two pieces of information are crucial in this respect:

1) Daily loading curve (DLC) of individual or aggregated end-users, which is mainly important for load forecasting and billing purposes. The volatility of the DLC, although very high at individual end-user's level, decreases with higher aggregation levels (see Figure 1.2 showing DLC at different aggregation levels based on actual data coming from a pilot site [24]);

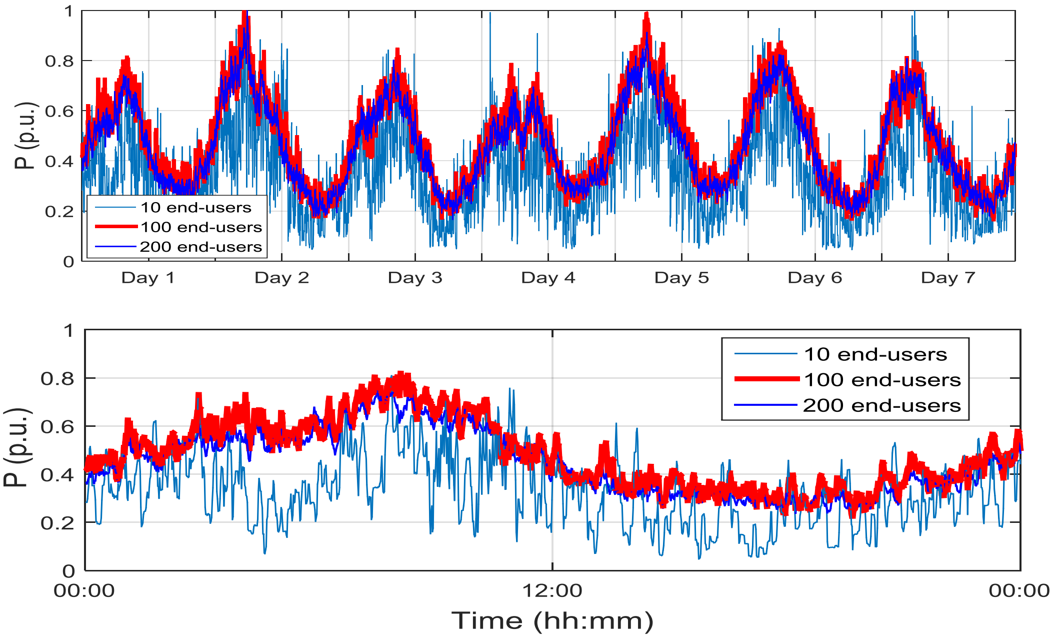


Figure 1.2 DLC during one week (top) and one first day of the week (bottom) at different aggregation levels

2) Composition of demand, i.e., the shares of different load types within the DLC, including both flexible (controllable) and inflexible loads.

1.5.1.1 Daily loading curve

In case of absence of real-time load measurements for individual consumers, it is useful to make an estimation of a typical DLC of the end-users. At this point, load patterns are reconstructed according to monthly energy consumption and typical load profile (TLP) of the end-users, i.e., the load class they belong to [25]. Load class profiles in the UK were introduced in 1994 in order to model involvement of different types of customers in the electricity market. This was supposed to save costs of installing half-hourly meters into every customer's premise. Eight classes of load profiles were established based on their annual consumption [26] and taking into account the peak load factor (LF), which is given as follows:

$$LF = \frac{Annual\ consumption\ (kWh)*100}{Maximum\ demand\ (kW)*Number\ of\ hours\ in\ the\ year} \tag{1.1}$$

Even when they may belong to the same type of activity or commercial code, the load patterns of consumers might be very different [26]. As [25] showed, there was a limited correlation between consumers' activity type (i.e., load class) and their load pattern. SMs should therefore be used to obtain more accurate load profiles from the end-users and link them to the appropriate class using some of the classification or clustering methods, which will be discussed in Sections 2.2.3 and 2.2.4 of the thesis, respectively. Classification/clustering of groups of customers is also necessary in cases where several retailers supply parts of the same feeder or a group of loads, so the load they supply individually can only be forecast by a bottom-up approach, i.e., from each customer [27]. When SM data are available, the typical load profile is calculated using the following steps:

1) Categorisation of measurements based on the season and type of the day (working day or weekend);

2) Normalisation of measurements according to the peak load of the consumer:

$$z_{ij} = \frac{x_{ij}}{maxX_j} \tag{1.2}$$

where z_{ij} and x_{ij} are normalized and real values, respectively, of the i-th element (time step) in the j-th consumer's load vector, while X_j is the load vector for the representative load pattern of the j-th consumer [28, 29]. In the case of the UK, where averaging period of a SM is 30 minutes (only the average value of the samples recorded over a 30 minute time window is reported), length of the vector X would be $24 \cdot 2\, samples = 48\, samples$. The reference power is the peak value of the average load pattern – therefore it does not correspond to the true peak power reached by the load pattern in the period of observation because of averaging [30]. Another approach for normalizing may be the max-min normalization formula, given in [31]:

$$l_{mh} = \frac{x_{mh} - min_{m=1,\dots,M}\{x_{mh}\}}{max_{m=1,\dots,M}\{x_{mh}\} - min_{m=1,\dots,M}\{x_{mh}\}} \tag{1.3}$$

where M is the number of patterns represented as vectors x_m $(m = 1, \dots, M)$, each containing H elements, i.e., time steps $(h = 1, \dots, H)$, and l_{mh} is the h-th element of the m-th pattern in the normalized dataset;

3) Smoothing, i.e., filtering choppiness of profiles due to random events or noise;
4) Cluster analysis (data mining method, which will be detailed in Chapter 2 of the thesis);
5) Determination of the TLP.

Apart from deterministic approaches, total aggregated load curve can be derived probabilistically based on limited statistical data from residential users – Monte Carlo simulations were used in [18] to generate the DLC of a group of users at the desired aggregation level. The authors argued that the probability distribution for characterising the aggregated daily load pattern depended on the time of the day and the level of aggregation. In addition, gamma and log-normal distribution [18] were recognised as the most suitable ones to probabilistically characterise the residential demand supplied by the same feeder or by the same substation.

Going further from assessing a DLC, even more challenging is to estimate impacts of different load categories (induction motors, lighting, resistive loads, power electronics, etc.) at the aggregation level. As the DLC of the end-users changes during the day, so does the load composition and the portion of controllable load. Therefore, the flexibility of the demand side varies in time, which is why the assessment (in real-time) or prediction (e.g., minutes or day-ahead) of the actual size of controllable loads can facilitate DR actions, as it can show whether the DR potential (load flexibility) is big enough for different needs of the DSO (reducing the cost of supply or obtaining reliability of the network). This way, timely assessment of the amount and type of controllable load facilitates load scheduling operations.

A formulation of aggregate demand flexibility based on collective behaviour of consumers was proposed in [32]. Flexibility is observed with respect to the probability to change the behaviour of aggregate users (either increase or decrease aggregate demand). The flexibility indicator is given as a percentage of the aggregate load that can be curtailed or increased without affecting the average change in aggregate demand. The analysis showed the importance of data granularity (i.e., averaging window) and aggregation level on the estimation of load flexibility. Higher granularity (smaller averaging window) may provide more detailed information about data flexibility, however, during low load periods it is affected by the operation of appliances with non-synchronous cycles (fridges, for example). At the same time, as the higher aggregation level smooths the loading curve (with fewer variations), it reduces the estimated flexibility indicator.

An example of industry practice for estimating TLP and DR potential is reported in [2], where historical load profiles of domestic, commercial and industrial users are clustered based on similarity. Customers belonging to clusters with low base consumption and high daily volatility are identified as those with high DR potential.

1.5.1.2 Load disaggregation

Load disaggregation methods refer to disaggregating (decomposing) the total consumption of an end-user, recorded by SMs, to individual appliance level, with the aim to reduce uncertainties in DSM programs [33]. These methods can be intrusive or non-intrusive. Intrusive methods involve measurements performed at each appliance circuit, for example using smart home plugs [34]. This approach requires investments into the installation of advanced monitoring (sub-metering) devices or smart home appliances, which are able to report their daily consumption with high granularity (e.g., every second). Installation of such devices may cost between $100 and $1,000 per household, as reported in [35]. It is expected that up to 30% of end-users in the UK will have smart wet appliances by 2030 [15].

Non-intrusive load monitoring (NILM) is based on pattern recognition of different electric appliances and it is performed at the customer supply point, for example at the cable supplying the end-user’s premise. Most commonly used electrical features or “signatures” for load disaggregation are active and reactive power (these differentiate between appliances with similar loads), electric harmonic and transient patterns [33]. Authors in [36] proposed an artificial neural network (ANN) – based approach for identifying individual appliances according to the current waveform measurements at the supply point. The multilayer perceptron (MLP) model (which will be detailed in Section 2.2.3.2 of the thesis) gave highly accurate results with samples measured in laboratory, but in order to deploy this type of solution practically, current waveform of each home appliance model would have to be monitored separately in order to record its harmonics in different operation modes. Methodology described in [37] used appliance consumption data measured every 10 seconds for training the Factorial Hidden Markov Model on a case study of 5 houses. However, the energy disaggregation results showed noticeable errors higher than 35% in most of the test cases.

The application of deep learning neural networks for load disaggregation was presented in [38]. Three architecture types of the neural network were investigated: *i)* long-short term memory, *ii)* denoising auto-encoders and *iii)* regress start time, end time and power. The method was used to disaggregate 5 types of home appliances, with 6-second sampling step. A separate network was trained for each appliance, where the training target was appliance consumption, and the input to the network was aggregate power demand. The results showed that in most cases the deep learning ANNs outperformed the benchmark methods (factorial hidden Markov model and combinatorial optimisation). However, the method was not validated on longer test data, but only on data from a few days.

A common drawback of the load signature methods is that they require a library of high-resolution measurements (usually faster than 1 Hz [39]) of appliance parameters needed to disaggregate the total load of an end-user, including current waveform, harmonics, switching transient waveform, etc., which are not always easily accessible, especially in case of a large number of users and at the scale of real distribution network communication systems.

Authors in [33] proposed a combination of smart metering technology with device-level load monitoring to disaggregate demand. The method proposes forming a database with recordings of real power load profiles of different appliances with two possible solutions: *i)* Generic load model – active load profile of a certain type of appliances, e.g., air conditioners; *ii)* Specific load model – active load profile of a specific model/brand of appliance. This way, customers would be able to choose from a library of profiles the one that fits their appliance best and enable the network operator to perform disaggregation based on the total load measurements. The application of this solution would not be trivial as there are numerous types of appliances on the market, especially the electronic ones, and having to update the database regularly would be a challenging task. Another problem might be that different appliances have similar real power pattern, and reactive power was not taken into account.

An approach for disaggregation of distribution feeder load in real time was suggested in [35]. The methodology relies on two on-line learning methods, namely dynamic mirror descent and dynamic fixed share, and aims at disaggregating total feeder load into heat, ventilation and air conditioning (HVAC) load and the rest of the load. Learning process is performed using historical data about real power measurements on the feeder, SM data (including sub-metering of HVAC loads) and outdoor temperature, while the real time data include feeder measurements and outdoor temperature. However, the suggested methodology did not propose an algorithm for disaggregating other load types (heaters, lighting, etc.). For example, in load disaggregation study presented in [40], apart from HVAC load, which was identified with relatively high accuracy (20% error), other appliances were disaggregated with much lower accuracy (errors higher than 90%).

1.5.1.3 Measurement-based and component-based load modelling

While load disaggregation is seen as a necessary step towards load profiling and more reliable DSM programs, load modelling has a significant impact on power network studies, including steady-state and dynamic analysis. Therefore, in order to assess the impact of the aggregate DSM on network performance, one needs to model the loads at network buses appropriately. Load modelling is a process of deriving parameters of a chosen load model, using either top-down or bottom-up approach, to represent load behaviour (static or dynamic) using mathematical models. Two basic groups of load models are static and dynamic. The former one models load irrespectively of time, while the latter one models load changes with time [41].

Some of the most frequently used static load models are exponential load model and the so called ZIP (polynomial) model [41]. The exponential model for real power (P) and reactive power (Q) is given with the following expressions:

$$P = P_0 \cdot \left(\frac{V}{V_0}\right)^{\mathrm{k}_P} \tag{1.4}$$

$$Q = Q_0 \cdot \left(\frac{V}{V_0}\right)^{\mathrm{k}_Q} \tag{1.5}$$

ZIP model, comprising three components, namely constant impedance, constant current and constant power load, can be formulated as follows:

$$P = P_0 \cdot \left(p_z \left(\frac{V}{V_0}\right)^2 + p_i \left(\frac{V}{V_0}\right) + p_p\right)(1 + F_P \Delta f) \tag{1.6}$$

$$Q = Q_0 \cdot \left(q_z \left(\frac{V}{V_0}\right)^2 + q_i \left(\frac{V}{V_0}\right) + q_p\right)(1 + F_Q \Delta f) \tag{1.7}$$

In the given expressions (1.4)-(1.7) V and V_0 are actual and initial (rated) voltage values, P_0 and Q_0 are initial values (at the rated voltage level) of the real and reactive load, and k_P and k_Q are voltage exponents of real and reactive power. p_z , p_i and p_p (or q_z , q_i and q_p in case of reactive power) are load participation indices corresponding to the three load components: constant impedance, current and power, respectively. At any point in time the sum of these three indices equals 1. F_P and F_Q are coefficients describing frequency dependence of loads, however, in most studies these are neglected due to small variation of frequency in most of the networks compared to variation in voltage. The ZIP load model is generally considered suitable for representing the modern non-linear loads [42]. Furthermore, for voltage stability analysis it is recommendable to include dynamic loads, i.e., induction motors (IMs), in the load model [43, 44]. This is typically done by presenting the equivalent composite load model as a parallel connection of ZIP load components and IM load [45].

Measurement based approach is a top-down approach deriving parameters of a chosen load model from system disturbance data using conventional and artificial intelligence based techniques and pattern recognition [46]. The derived model is then validated by comparing the simulated response with the measured one. As the process depends on measurements of the changes in load, one should distinguish between load changes due to grid events (e.g., sudden voltage drops or frequency change) and natural daily changes in load. One way to overcome this issue is using fuzzy inference system approach [47].

Component-based approach is a bottom-up approach which derives the overall load models by aggregating (i.e., summing up with corresponding weighting factors based on participation of different load categories in total demand) known individual load models within corresponding load sectors/classes at the bulk supply point [46]. The key information required for estimating dynamic response of demand at the bulk supply point (following a disturbance) is the type and the percentage of different load categories, rather than appliance or end-users involved [46]. The most comprehensive overview of load models, load modelling approaches and the effect of load models on system performance can be found in [48].

1.5.1.4 Probabilistic approaches to decomposition of aggregated demand

In the absence of real measurements, probabilistic approaches are taken to model demand composition. For example, Markov chain Monte Carlo approach was used in [42] to derive individual residential load profiles (both real and reactive power) in residential sector based on statistical data for the UK. Furthermore, a load model at the aggregation level (10,000 customers) was developed based on the shares of different load categories within the total demand. The following load categories were identified: power electronics, resistive loads (heaters), lighting, directly connected motors (white appliances and water pumps) and drive controlled motors (HVAC). The proposed model gave very high confidence (absolute percentage error up to 5%) in estimating aggregate consumption of individual load categories, using UK-wide statistical data as the base case. The accuracy of the approach, however, was not investigated at lower aggregation levels, where demand tends to be more volatile, nor was it compared against actual measurements.

As another approach to decomposition of demand in the presence of demand side uncertainties, probabilistic load disaggregation into load categories was developed in [16] based on the measurements of total active and reactive load at a bulk point using ANN and statistical data about typical load composition in residential, commercial and industrial sectors. This work proved the validity of the approach showing reasonably good accuracy in the estimation of load composition (the 95th percentile of the absolute relative error of estimating shares of different load categories was between 0 and 10%). At the same time, both training and testing data were generated in a probabilistic manner, using randomization of voltage at primary substation and participation of load categories. Therefore, further validation and adjustment of the approach is required with incorporation of more realistic consumption data reflecting data streams coming from numerous SMs.

1.5.2 Demand response programs

DR is a common name for changes in load consumption (increase/decrease) as a response to external signals, motivated by either environmental, market or network implications [49]. DR has been recognized as one of potentially cost-effective options for operating the power network [50]. Typical aims of DR are maximising the use of renewables, maximising the economic benefit, minimising the energy import from the main grid, or reducing peak demand [51]. Large-scale (aggregate) DSM can be used to provide balancing services by selling flexibility [52], compensate RES volatility [53], provide regulation services [10], minimise network losses or defer investment and contribute to network security and reliability [54-56].

The main categories of DR programs are based on the type of the external signal motivating the end-users to change their consumption, and given in two groups, as follows [57-59]:

1) Price-based DR, divided into:
    - Time-of-use (ToU);
    - Real-time pricing (RTP);
    - Critical peak pricing (CPP);
2) Incentive-based DR, with its most common forms:
    - Direct load control;
    - Interruptible load contract (ILC);
    - Demand-side bidding (DSB);

  - Peak time rebate (PTR).

Price-based DR programs are relying on different tariffs for the end-users, either depending on the period of the day (ToU pricing), following the real change of wholesale electricity price on the market every hour or 10-15 minutes (RTP), or increasing the price substantially during very high (peak) loading periods – either in predefined periods or with a few hours notification (CPP). In CPP programs, the number of critical days per year is predefined, but the timing is unknown, the notice being usually one day [55]. The EDF in France, for example, has 10 million customers on this type of program.

In the direct load control scenario, the system operator has direct control over the demand covered by the DR program [60]. In this respect, load can be dispatchable (i.e., it can be curtailed according to a signal from the system operator) or non-dispatchable (i.e., curtailed manually, by the customers, following an incentive). Automated DR (with dispatchable loads) enables predictability and persistence of end-users' response over longer periods [61]. Control relays which interrupt power to loads are usually activated via radio signal, telephone, or using the power lines [62]. Typically reported applications of direct load control include voltage control, provision of ancillary services and energy arbitrage [9].

Interruptible programs are designed for large customers who have to offer significant reduction (e.g., of at least 1 MW) and are ready to execute it at any time [55]. The minimum notification time (usually 10 minutes to one hour), maximum interruption duration and maximum number of interruptions per year are pre-defined. Communication is commonly done by phone, email or fax. Larger industrial users usually have a back-up generator which is turned on to prevent production losses during load curtailment. Demand-side bidding represents participation of load flexibility in electricity market. In PTR contracts, the consumers agree to reduce consumption during peak pricing times and receive a rebate in return.

1.5.2.1 Incentivising end-users

According to [11], the most common methods to incentivise DR are ToU tariffs and direct load control. Apart from financial incentives, some end-users may be interested in ecological aspects of DR, i.e., the reduction of CO_2 emission. It has been reported that some companies participate in DR in order to improve their brand image [61].

Environmentally driven DR aims to improve the environmental and social standards by reducing energy use through energy efficiency programmes and by reducing greenhouse gas emissions [49]. Finally, different groups of end-users will react to different incentives, which is why some authors [63] suggest tailoring types of incentives based on motivation of the end-users, for example sending climate information messages to those who give priority to ecology, or financial savings for those with self-interest.

End-users can be incentivised for efficiency or curtailment behaviour [64]. Efficiency behaviour refers to capital (one-off) investment for reducing fossil-energy use, e.g., home retrofitting. Curtailment behaviour relates to repetitive reduction of fossil-energy use, by changing daily consumption patterns. Social scientists have investigated different approaches to change customers' behaviour with respect to electricity usage and sustainability [7]. They have suggested different solutions, for example: *i)* visualisation (e.g., using thermal imaging to show indoor heat losses to encourage home retrofitting), *ii)* tailored information about energy-saving measures and the impact these could have on customers' electricity bills, or *iii)* social comparison, where customers are informed about their electricity usage compared to their neighbours, for instance. Report [64] indicated that monetary incentive alone (in this example, €1.7 for each 1% reduction in energy use) resulted in lower consumption reduction than when it was combined with social comparison. The former program resulted in 5.9% reduction of energy use, while the latter one resulted in 8.2% reduction.

Some drawbacks of the monetary incentives have been recognised in [64]:

- Rebound effect (savings in electricity bill may induce investments in other products that require increased electricity use);
- Low individual benefit for the consumers, i.e., low cost-effectiveness of participation in DSM from the end-user's perspective;
- Undermining intrinsic motivation that is based on environmental benefits: a research has shown that emphasis on environmental benefits alone attracted more participants in DSM than emphasis on monetary benefits only or these two benefits in combination, as shown in Figure 1.3. Similarly, social reward (e.g.,

recognition) also resulted in higher reduction of energy use compared to monetary incentives.

Finally, it can be concluded that socio-economic profiling of the end-users in an area is a necessary step before introducing different types of incentives for DSM participation.

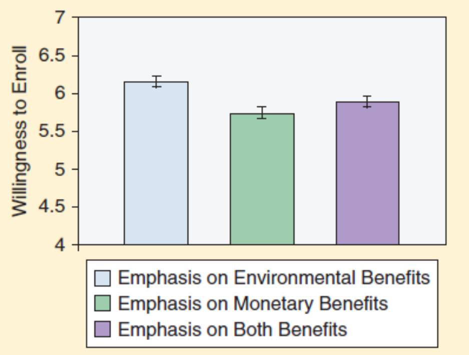


Figure 1.3 Willingness to participate in DSM program based on different incentives (adopted from [64])

1.5.2.2 Demand response approaches

Authors in [65] investigated a game theory-based DR for price-responsive appliances. The individual appliances schedule their consumption at the times of lowest electricity price with the objective of minimising the operating cost. In this distributed approach, the price signal is sequentially sent from a centralised entity to each appliance, which then performs power change to reduce its energy cost. The proposed methodology avoids new demand peaks and successfully flattens the total loading curve of the aggregated demand. An incentive-based approach was examined in [59], where the end-users' elasticity is encouraged by voluntary coupon incentives when the wholesale real-time price (paid by the retailer) exceeds the fixed retail price (paid by the end-users). This idea was motivated by the overbooking strategy of airline companies, as the end-users are expected to reduce their consumption as long as the revenue they receive for that exceeds the benefit they would get if they did not reduce the consumption. In other words, participating end-users can only save money. The main contribution of the approach is that significant benefits are achieved even with moderate participation of the end-users. The difference between this program and PTR is that in the former program the rebate rate (i.e., the coupon price) is not fixed, but updated iteratively between the retailer and the customers, based on the close to real-time system conditions.

Authors in [57] are using demand-side bidding model and propose a two-level optimisation – at the upper level, the grid operator minimises the cost of DR by optimising load shifting schedule based on a load shifting bidding curve at each load bus. Then, at the bottom (nodal) level, the demand of each node is rescheduled based on the requirements coming from the upper level.

Optimal load scheduling is performed in [66] to minimise the total energy consumption cost. The problem is constrained by the capacity limit of the distribution transformer, and the delay quality of service for different demand blocks. The scheduling process is divided into two steps: capacity planning and real-time scheduling. Similar approach was taken in [67], where the individual appliance scheduling was done in two steps: first day-ahead, following the hourly price forecast and expected users' behaviour, while the second optimisation step was performed in real time (in minutes), following the actual prices and users' comfort.

As the high penetration of DER in some areas may have effect on sub-transmission and transmission levels, TSOs are advised to control aggregated DERs [68]. Similarly to DGs and storage systems, flexible (controllable) loads can also be considered as DERs. Following this, authors in [69] presented an energy management system as interaction between the DSO, aggregators (clusters of the same type of DER connected to one bus) and dispatching centre of the upper grid. In this scenario, DSO first receives information from aggregators about available (forecast) flexibility and information from upper grid dispatchers' needs, and then makes the optimal schedule and sends it back to the other two actors.

Another hierarchical DR program was seen in [70], where demand flexibility was observed as virtual state of charge (SoC) of aggregated thermostatically controlled loads (TCLs) modelled as virtual generators (VG) with negative output. The main aim of DR in this example is providing balancing services to the upstream network by following a target load curve. The SoC is proposed to define the upward and downward regulation capacity of the load group. TCL is defined by a state vector consisting of indoor temperature, on/off state and corresponding power. TCLs (heat pumps) are controlled sequentially using a state-queueing model, where the control signal (ON/OFF) is sent to the units prioritised based on the indoor temperature. The central controller, which is equipped with indoor temperature forecaster for the next time step, determines the ON/OFF status of the HVAC units and creates the priority list. The feasible (flexible) region of the VG is determined by aggregating rated (real) power of ON and OFF devices and adding them to or subtracting them from, respectively, the total power consumption without control (the consumption that existed before any of the heat pumps was controlled). It is assumed that the state of each heat pump can be monitored. In cases where there are multiple VGs in the network area, these can be aggregated to virtual power plants (VPP). The flexibility boundaries of VPP would be obtained by aggregating boundaries of individual VGs.

Based on the observed literature examples, typical DR architectures are centralised ([60, 62]), hierarchical ([62, 69, 70]), and distributed ([59, 65]). A typical hierarchical architecture, adopted in this thesis, was proposed by [71], where the main actors are the TSO (the DR initiator), the aggregator (managing and selling flexibility from aggregate end-users), the DSO (monitoring DR action in order to preserve its operating conditions, namely voltages and currents), and the end-users willing to participate in DR programs.

In spite of potentially significant benefits that could arise from effective DSM, there are some notable obstacles for harnessing flexibility from the demand side. These include [30, 62]:

- Lack of controllable loads or infrastructure needed for automatic load control;
- Demand uncertainties, which are more emphasized at lower aggregation levels;
- Thermal inertia affecting the scheduling of thermostatically controlled loads (TCLs);
- Load (energy) payback, which, if not properly planned, may result in peak loading at time steps following load curtailment. Load payback is often modelled in a simplified manner, as a redistribution of curtailed energy during the control period;
- Customers’ lifestyle and comfort, i.e., their willingness to participate in DR programs;
- Inter-temporal constraints (e.g., storage capacity, ramp rates);
- Costs for the DR responsible entity on one hand and limited revenues for the end-users on the other.

1.5.2.3 Demand response and load composition

Unlike conventional generators, DR (similar to energy storage) has limited duration of response, hence, load reductions have to be scheduled to provide energy services during periods of highest value [72]. Furthermore, the aggregate response is more predictable and reliable than the one of individual end-users, especially if the end-users from the same demand sector are aggregated. Aggregation allows longer curtailments by sequential shedding of individual loads within the aggregated group, and varying response times from individual loads within the group [73]. Therefore, there is no need to model in detail operating characteristics of different types of load in the aggregation. As the communication technologies are already in place, including the smart metering systems, it could be argued that the only obstacle to reliable aggregated DR programs are appropriate load models and control strategies [62].

Different approaches to modelling aggregate demand flexibility have been suggested in the literature:

- The use of sensitivity functions indicating each user’s probability of shifting usage of each device by a certain time, given the reward in the new period of usage [74];
- The unit commitment optimisation approach, to compare flexibility from demand-side resources with the one from fast ramping generation [62, 75, 76];
- Probabilistic demand curve, similar to generation availability curve of a renewable energy source (RES) [77];
- Storage model, where demand flexibility is observed as virtual state of charge (SoC) [70]. Similarly, load availability is often given as upward or downward flexibility, referring to load decrease or increase, respectively [10, 78];
- Load shifting bidding curve showing the maximum change demand is willing to make based on the price [57, 79].

Information about load flexibility is beneficial for the assessment of DR potential and the outcome of a DR program. Furthermore, information about the composition of demand, with respect to different load categories, informs the DR responsible party about the shares of different types of load that may have different availability for DR and bring different benefits. For example, a case study examined in [55] showed that participation of one load category (washing machines) in DR brought more financial benefits than participation of another one (in this case, water heaters). In [80], end-users receive a reward for their load shifting participation, as well as for the voltage improvement in the supply feeder. Following this approach, prediction of available demand flexibility, i.e., short-term forecast of load composition, for example day-ahead, allows the supplier to adjust the level of incentives in case there is a need for attracting more end-users to participate in DR actions. In direct load control programs, described in [50], load composition could enable the system operator to have a more accurate overview of the size of available controllable loads whose consumption can be remotely controlled.

Another type of DR that can benefit from information on load composition is voltage-based DR (conservation voltage reduction or CVR), which is a non-intrusive DR program, as it does not affect the end-users’ comfort. It has been proven in [81] that in this case demand regulation potential (i.e., demand flexibility) depends on the initial size of load, as well as on load composition, namely the participation of load that can be modelled as constant impedance. Similarly, methodology for assessment of voltage-based DR potential in UK, described in [82], raised the need for information on the load composition at distribution system buses. This way, the operator can estimate with higher accuracy what the available load flexibility coming from this type of DR will be and whether additional actions are necessary to meet desired aims.

Load categorisation for DR was performed in [83]. Residential loads were classified into three groups: automatically controlled appliances with large demand and one run per day (storage heaters), automatically controlled smaller appliances with frequent runs during the day (fridges) and semi-automatically controlled appliances operating few times a week (e.g., washing machines). The authors showed that load control of EVs, batteries and storage heaters could bring the highest DR revenues.

Identification of controllable HVAC devices and optimal scheduling of these loads to meet the aggregate target loading curve were performed in [84, 85], as part of a direct load control program. Decomposition was performed using a NILM method applied to current subharmonic waveforms of the loads. The target curve is predefined based on peak clipping and valley filling, which implies that the total energy consumed during the control period should be the same as the one without any DR program. Meeting the target loading curve was accomplished using multivariable predictive control, namely sequential quadratic programming, of HVAC groups. The optimisation takes into account load payback of each controlled group, and limits it by a predefined value. The methodology, however, did not consider other controllable load groups.

Demand composition, as reported in previous examples, is highly important for assessing the available demand flexibility for these programs. On the other hand, curtailing or shifting demand during DR actions may change load composition, and thereby affect the dynamic response of aggregated demand in case of a network disturbance (e.g., a voltage step change). Dynamic response represents change in active/reactive power of the load following a step change in voltage, which might affect the angular and voltage stability of the power system [39]. The size and shape of the dynamic response mainly depends on the size and composition of the load, which is why load decomposition finds its application in this area of power system studies.

It has been shown that induction motors, thermostatic loads and energy efficient devices present sources of load dynamics [17], therefore their effect on the aggregated load demand response should be further analysed. Special attention should be given to new types of load, mostly non-linear power electronic devices (DC power supply loads, light emitting diode (LED) and drive-controlled motor loads). These types of load need to be modelled using dynamic load models, while resistive loads, common lighting and similar can be represented using static models (e.g., exponential or polynomial) [48]. Reference [47] examined application of a functional polynomial network (FPN) for clustering load responses to voltage and frequency disturbances based on different load composition. A load model would then be created for each cluster. The authors proved that the use of linear and ZIP load models was not justified in cases of dynamic responses. Instead, they used machine learning system (the FPN) trained by real measurements to provide more accurate assessment of the dynamic response following a voltage and/or frequency change. Finally, knowing the effect of different load compositions on system stability, one may take measures, i.e., appropriate DR actions, to prevent those compositions of load that might provoke instabilities in the power system. Dynamic response of demand-based load shifting as a part of DSM program was introduced in [86].

1.5.2.4 Demand response and network performance indicators

DR programs including large users or aggregations of users visibly change power flows in the network and thus may affect different indicators of network performance. Recent literature has investigated how DR can be used to support voltage and frequency control during regular or disturbed operating conditions. Another area or research has been the effect of DR on network losses.

1.5.2.4.1 Voltage support

Voltage stability is described as the ability of the system to maintain acceptable voltages at all buses during normal operating conditions and after a disturbance [44]. Voltage instability occurs when a disturbance, a load increase or a change in system conditions cause voltage to drop progressively. Voltage stability is assessed through different indicators. One of the most common indices is the calculation of the minimum singular value of the power flow Jacobian matrix [87]. The minimum singular value represents the distance between the current operating point and the singularity of the power flow Jacobian matrix [88]. When the Jacobian matrix is singular there is no inverse matrix, i.e., there is an infinite sensitivity of the power flow solution to the small changes in parameters, and the power flow solution cannot be obtained. In modal analysis [44], eigenvectors and eigenvalues of the reduced Jacobian matrix are calculated – this approach is very useful in determining the critical elements in critical areas in the network with respect to voltage stability.

Except for the modal analysis, a typical approach for assessing voltage stability is via network load margin. For a certain operating point, the load margin is the amount of additional load in the network that would cause a voltage collapse [89]. During a heavily loaded condition, even a relatively small, but sudden increase in demand can result in voltage instability. The load margin is commonly derived from the real power–voltage characteristic, the so called PV (or “nose”) curve, depicted in Figure 1.4. The PV curve will change due to contingencies, resulting in lower load margin even at the same operating point (OP). If the current OP is at the upper half of the PV curve, the real power margin is the amount of load increase that will cause the power system to reach the maximum loading point (P_M' in Figure 1.4). The change in network parameters (due to a contingency, for example) changes the PV characteristic (dashed curve in the figure), while the change in load (constant power load characteristic is used in Figure 1.4 to illustrate the concept) changes the position of the OP on the PV curve, and consequently the load margin (the distance between OP and P'_M in Figure 1.4). Changing the network parameters, by adding FACTS devices, for example [90], moves the PV characteristic, while changing the load by DSM action can move the OP to left or to the right (along the PV characteristic), and such increase or decrease its distance to the maximum loading point, respectively.

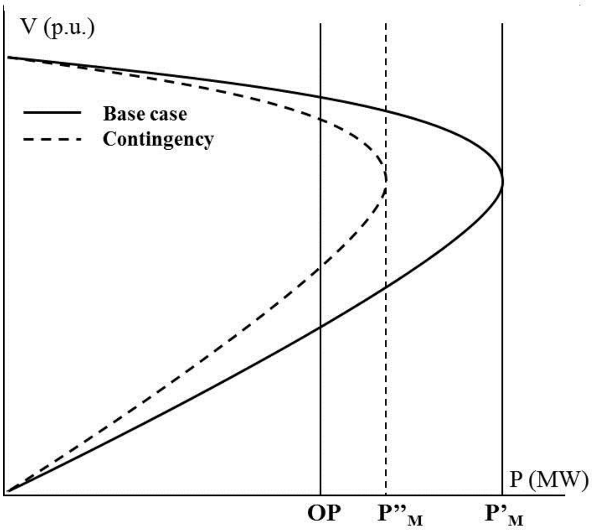


Figure 1.4 PV curve (adapted from [91])

Keeping the load margin as large as possible ensures that the system will be able to withstand disturbances and unexpected increase in the load without endangering its voltage stability. Typical method for deriving the load margin is continuation power flow, which, unlike the Newton-Raphson load flow method, permits convergence around the saddle node (tip of the PV curve) [92]. The method uses constant power load model and allows for obtaining the load margin and identifying the weakest bus in the network. The sensitivity of the bus to load increase is detected by following the ratio (dV_{bus}/dP_{Total}), where dV_{bus} is the change in bus voltage with the change in total load (dP_{Total}) of the system. The location of the weakest bus may change with the changes in size, characteristics, and location of the load. The sensitivity (dV_{bus}/dP_{Totall}) will be negative if the loading margin is in the stable zone, and close to zero if it is far from the maximum loading point.

Another voltage stability indicator, the L-index, uses reconfigured admittance matrix to assign a value from 0 to 1 to each load bus [12] - the higher the value is, the closer the bus and the system are to the voltage instability point. Other indices include: sensitivity of impedance ratio index, V-Q sensitivity index, index i, voltage collapse index (Lambda), channel components transform, diagonal element dependent index, etc. [13].

An important part of accurate voltage stability analysis is appropriate load modelling [44] – for example, at lower network voltages (i.e., below 85-90% of the nominal value), some induction motors may stall and draw more reactive power, causing this way the voltages to drop even more. In the long-term voltage stability analysis performed in [43] distribution network loads were modelled using a combination of voltage sensitive and induction motor loads.

Even though voltage stability phenomena have been widely analysed at the transmission network level, voltage stability in distribution networks and microgrids has been given more attention recently. Key factors in voltage instability in microgrids are DER limits and sensitivity of load to voltage [23]. Even though voltage collapse cannot be observed in microgrids, instabilities may be seen in the form of unacceptably low steady-state and dynamic voltages [23]. Furthermore, clogging (or “radial”) voltage instability happens in distribution, sub-transmission, and occasionally transmission network [19]. It occurs due to series reactive losses, on load tap changers (OLTCs) reaching tap limits or shunt capacitors reaching susceptance limits. At the same time, no support in reactive reserves appears in generators, static VAR compensators or synchronous condensers. These can “clog” the network and prevent reactive power flow needed to support voltage drop in sub-areas of the network. This instability, caused by increased transfer, can be assessed by loadability assessment methods or PV curve.

Distribution network voltages have traditionally been regulated using transformers with OLTCs [43], however, with the increased loading of distribution networks (coming from proliferation of both new types of load and distributed renewable generation), other types of resources may be needed to complement the existing ones. This is of particular importance in degraded operating conditions of the transmission network, where transmission network voltages are not as stiff as expected, and may need more support from the distribution network [43].

Previous studies have dealt, to an extent, with the influence of DGs and DSM on voltage stability. Location of the DGs in the network affects the voltage profile, which is why optimal allocation of these resources is very important for appropriate voltage support. Candidate buses for installation of DGs, prioritised based on their sensitivity to voltage profile and thus capability to improve the voltage stability margin was discussed in [20]. Voltage stability was assessed with the maximum loadability of the network, i.e., the load margin. Mixed-integer nonlinear programming was applied for this purpose, with an objective function of improving the stability margin. The constraints were the system voltage limits, capacity of the feeders, and the distributed generation (DG) penetration level. In another example [22], modal analysis and continuation power flow were used to determine optimal locations of DGs in distribution network. The placement was then evaluated by assessing voltage load margin, active and reactive loss reduction and voltage deviation over all network buses. In both DG allocation problems the methodology was illustrated using a radial distribution network, namely the IEEE 41 bus network [20] and IEEE 33 bus network [22],\ while demand was modelled using the constant power load model.

In [88] DR was used in contingency events, to support voltage stability until reserve DGs get connected. Due to the ramp limits of the generators, the approach proposes fast responsive flexible loads to support the voltage stability only until the generators get fully connected. After a disturbance, it is assumed that the load changes – increases in some buses, and decreases in other, keeping the total flexible load constant, until the old and additional generators are re-dispatched. Once the generators are dispatched, the loads return to their normal consumption plus/minus the load payback from the DR period. It was assumed that 100% of the demand was flexible, and that load could be completely curtailed at some buses, and increased (by the same amount) at others, to maintain the frequency. This approach, however, would be hardly feasible to deploy in reality, due to the limitations in load flexibility (both upwards and downwards). Authors in [49] used a multi-objective optimisation to allocate a limited number of network buses for provision of DR based on Pareto optimal solution, i.e., the solution whose improvement in one of the objective functions would deteriorate at least one more function from the given set of objective functions. The objective functions included: generation scheduling cost, voltage drop, voltage stability margin, network loss, and incentive payment while, crucially, the demand-side flexibility was assumed, but not clearly evaluated, and the effect of load payback was not accounted for. In both of these studies [49, 88] the load was modelled as constant power, though the most unfavourable for voltage stability, as it usually does not reflect the behaviour of the actual load in the network. Authors in [93] considered DSM as an alternative to short-term voltage stability improvement. In this study the load was modelled as a composition of different static and dynamic loads, but the criteria for curtailment or shifting different load types was not defined. Similarly, [54] examined how DSM can affect the estimated voltage stability margin by considering different load models, but fell short of performing optimal allocation of DSM for improving voltage stability.

Optimal DR was presented in [76] where both network requirements (preserved voltage stability) and constraints based on end-users' comfort are met. The approach relies on highly distributed heat pumps whose aggregated output can be ramped up or down from a centralised controller behaving as a virtual generator, similarly to [70]. Security constrained OPF is performed with the end-users' constraints included. Voltage stability margin was observed at initial loading conditions (before any control actions) and then, in the periods where it was lower than the critical value, DR was triggered. The DR is performed via the central controller (power system operator) who fetches flexibility boundaries from aggregated groups of heat pumps and reschedules their usage. The flexibility boundaries are imposed by the temperature comfort of the end-users, so the central controller can only define a new scheduling target, for solving voltage instability issues, within these boundaries. The objective of the OPF was minimisation of the control cost, constrained by the load margin (which has to be higher than the critical one) and the flexibility region of the virtual generator. The load margin-induced constraint is incorporated using the sensitivity of the load margin to load bus injection at the critical (voltage collapse) point following a contingency. The OPF problem is solved iteratively, until there are no contingencies recorded, and the solution is given as a vector of power injections at each controllable load bus. Although the approach successfully tackled voltage instability issues, it relied only on one load type. If other types of load were considered, payback load would have to be modelled as well, in order to analyse possible effect of load increase following a load decrease. Furthermore, load model used in the study was not defined.

1.5.2.4.2 Frequency support

Frequency is seen as a measure of real power balance between generation and load and thus should be kept constant (within a predefined range) for normal operation of the power system. There are three levels of frequency control in practice: *i)* primary control (provided by frequency containment reserves) is a fast local automatic control adjusting the real power generation and consumption to quickly restore the balance between generation and load, *ii)* secondary or load-frequency control (provided by frequency restoration reserves) is a centralized automatic control adjusting the real power production or consumption to restore the frequency and the interchanges with other systems to their target values, and *iii)* tertiary control, which dispatches the generators, bringing the power flows back to their target values [94]. Primary control should rely on resources distributed across the network to avoid large unplanned power transfers following a disturbance [95]. Therefore, distributed DSM resources, with the self-regulating effect of frequency sensitive loads (motors) or frequency sensitive relays, can be valuable for this type of frequency control in particular, complementing the speed governors of the generator units. Authors in [95] assumed that generators were able to provide 7% of their nominal capacity for primary frequency response, and that 10% of the load was frequency responsive. Participation of the demand drastically reduces the amount of response required by the generators.

Demand-side contribution to primary frequency control was analysed in [96], following the fact that individual generator’s response depends on its droop characteristic and local frequency measurement, not on a signal from a control centre (this type of signal is usually sent for secondary or tertiary control loops). Therefore, for primary frequency control, there is no need for generator (or load) to be connected to a communication system to participate in frequency response. The proposed appliances participating in frequency response are the energy consumers (not power consumers) as they can be shifted in time, as long as they consume the predefined amount of energy (e.g., fridges/freezers, HVAC units, tumble dryers, water and space heating). This paper analysed bounds on the amount of frequency-sensitive (flexible) demand response which could be achieved in a power system. In another approach [94], primary and secondary frequency support by aggregated EVs and water heaters was examined. The DSM program is designed with an aim of providing the requested services with maximised social welfare of the end-users. A multi-agent framework shifts demand in time and provides the primary or secondary frequency support using the available reserve. In this price-based framework, the electricity price reflects the changes in frequency. Therefore, the controllable devices adjust their consumption based on the price-frequency dependency.

As the frequency response of demand always remains uncertain, the operator will need information about the aggregated frequency-sensitive demand response characteristic. The results in [94, 96] demonstrate that the aggregated active load response characteristic is similar to the droop characteristic of a generating unit. The aggregated demand can mimic behaviour of a generator in this respect, however the recommendation is to use demand as a complement to generation reserve, as otherwise the aggregation would need to contain a significant number of loads. Fast response of demand can be used to reduce frequency drop during a disturbance before conventional generators start restoring the frequency.

DR for both voltage and frequency support was developed in [97], where all responsive devices were classified based on their controllability degree. Once a violation of voltage or frequency is detected, the control signal is sent to the most influential (i.e., most sensitive) buses to change their active or reactive load. In the proposed scenario, hierarchical system is established where a central energy management system sends requirements for corrective actions to transmission agents (TAs) who calculate the requested active/reactive power change and forward it to the distribution agents (DAs). Finally, DAs send requests for active and reactive power change to the controllable loads based on the request from the TAs, available demand controllability, and distribution network voltage and line flow constraints. The multi-objective problem (minimising frequency and voltage deviations, and minimising manipulated active and reactive power to meet the overall goal) was solved using particle swarm optimisation (PSO). While the optimisation decides on the changes in the amount of active and reactive power, it does not consider the types of loads that participate in these changes. All the load buses are modelled in the same way, using composite ZIP load model with frequency dependence with constant load model parameters. The daily changes in load composition, and subsequently, in the parameters of the equivalent load model, however, are not accounted for in the analysis, only the time-varying controllability is considered.

1.5.2.4.3 Distribution network losses

Distribution network losses, i.e., the difference in electrical power delivered to the distribution network and the power delivered to the end-users, play an important part in the overall operational costs of a DNO, as they are affecting the carbon emissions and generator capacity requirements [98]. It was suggested in [71] that power losses, voltage dependency of the load buses and load payback should be considered in DR programs. Quantitative analysis of losses in the UK distribution network were provided in [98], reporting 1.5% losses at 33 kV level, and 3% losses at 11 kV level.

DR for minimisation of cost for the end-users and distribution network losses was performed in [99] using a two-level optimisation. At the lower level sequential quadratic programming (SQP) is applied - loads are shifted, based on the given daily electricity price, from periods of higher price to the periods of lower price. At the higher level, PSO is used to optimise the daily price that will, indirectly, through load shifting, minimise the network losses. For each generated swarm (a “candidate” daily price vector) of the PSO algorithm loads are scheduled using the SQP method to minimise the overall cost, and the consequent network losses are calculated. Finally, the result of the PSO algorithm is the price vector giving minimal losses. The distribution network losses were reduced by 12% compared to the base case without DR and with constant electricity price. The authors considered participation of residential, commercial and industrial loads, however they did not distinguish between different load models representing these load sectors. In addition, the power factor of each load bus was considered constant even after load shift.

Minimisation of network losses and generation cost based on real-time scheduling of EVs was suggested in [100]. Losses are minimised by prioritising charging of the EVs causing minimum impact on the network losses. The EVs are prioritised based on the sensitivity analysis of system losses to small variations in EV charging load at a given time step. These sensitivities are calculated from the Jacobian matrix of the power flow - this approach is called the maximum sensitivities selection. All loads were modelled as constant power loads.

1.5.3 Ancillary services provided by large-scale demand response

Ancillary services represent network services, provided to the network operator by different actors in a deregulated power system, which have two main goals [101]: *i)* maintaining a constant balance between generation and load, and *ii)* managing power flows within the network constraints. The main resources for ancillary services used to be conventional generators, which have recently been complemented by DER (including DGs, storage and DSM). Ancillary services can be classified into the following groups [101]:

- Continuous (frequency) regulation: provided by resources with automatic control for minute-to-minute balancing between generation and load;
- Energy imbalance management (load following): slower than continuous regulation, bridging regulation service and hour-to-hour or half hourly bid-in energy schedules. This mechanism allows market clearing;
- Instantaneous contingency reserves: provided by sources that have frequency or other type of control that can rapidly increase output or decrease consumption as a response to a disturbance;
- Replacement reserves: provided by resources with slower response that can replace or complement instantaneous contingency reserve;
- Voltage control: injection or absorption of reactive power for maintaining transmission system voltages;
- Black start: generation units capable of starting themselves without any support from the grid, with sufficient active and reactive power to be used in system restoration.

The first four groups fall into operating reserves which can be further distinguished as either spinning (connected and synchronised with the system) or non-spinning (available and ready to be connected and synchronised within 10-30 min) [72]. Frequency regulation always comes from spinning resources, but the rest of operating reserves can come from a combination of spinning and non-spinning resources.

DR, as one of the new ancillary services providers, is expected to provide energy services (shedding or shifting load) and operating reserve (frequency regulation due to unpredicted short-term changes in net load/RES generation, contingency reserve following a fault, or flexibility/ramping reserve at times of large and unexpected RES ramp events) [102]. Energy reserves are sold in kWh (MWh), while operating reserves are sold in MW during a particular time period [72].

Distribution network can also be seen as one of the providers of some of the aforementioned ancillary services at the grid supply point (GSP). End-users with automatic control may have faster response than the conventional generators as the latter ones have slower ramping characteristics due to the inertia of thermal and mechanical systems [103]. It was reported in [62] that the ramp rate of loads is often constrained only by the speed of the communication network. Also, control of a large number of smaller units is less risky than controlling a small number of large units. DR resources could provide ancillary services at a lower cost and with lower carbon footprint than conventional generators [104]. There are 225 GSPs (mostly 400 or 275 kV to 132 or 66 kV) in England and Wales and another 280 in Scotland [98]. This gives around 500 points across the transmission network at which the DSOs (i.e., the demand side) could potentially be providing ancillary services through DR programs and other types of control (e.g., CVR).

As an example, Customer Load Active System Services (CLASS) project [105] was run by Electricity North West (one of the DNOs in UK) and National Grid (TSO in UK) to investigate and better understand the effects of network voltage change on electricity demand. The project demonstrated that through the deployment of voltage control equipment at DNO substations, network transformers could be used to modify network voltage and demand and so provide frequency and voltage management services to National Grid. The project identified that, if the voltage control techniques were applied widely, around 3 GW of demand reduction or demand increase could be achieved to provide frequency services, and around 2 Gvar of reactive demand could be achieved to help manage transmission voltage.

The distribution network is expected to provide ancillary services to the transmission network using flexibility of the load and distributed generation (DG). In that respect, controllable loads (that can be controlled remotely by the system operator through direct load control programs) and dispatchable DG sources could participate in ancillary services market [106]. Controllable loads could be used for generation-demand balancing, frequency control, peak reduction and network congestion mitigation [107]. Following approach in [97], load buses could be classified based on the type of ancillary services they can provide as a support to generators: *i)* voltage control would be obtained from buses providing real power support and those providing reactive power support (these buses have higher sensitivity of voltage to changes in real/reactive demand); *ii)* frequency control could be obtained by all system buses participating in real power support.

Potential benefits of loads participating in ancillary services are as follows [62, 101, 102]:

- Improved system reliability (more sources, which are distributed across the network, and are thus capable of providing spatially precise responses to contingencies);
- Improved market efficiency (lower price with more participants):
    - as the energy services provider, DR reduces the use of highest cost generators;
    - as reserves provider, DR reduces the use of less efficient partially loaded thermal units, as well as the variable cost associated to conventional generation providing regulation services;
- Market power mitigation (preventing generators from bidding up the price of ancillary services);
- Improved system efficiency and planning (avoiding uneconomical operations);
- Improved risk management.

Furthermore, provision of ancillary services is defined by the following deployment times [95]:

- Deployment start – maximum time between the TSO’s request and the start of the response;
- Full availability – maximum time between receiving the request and delivering full response;
- Deployment end – maximum time during which the service must be provided starting from the time of the request.

Ofgem (The Office of Gas and Electricity Markets) in the UK defined reliability rules and market design as source-neutral, allowing equal participation of load and generators. PJM (Pennsylvania, New Jersey, Maryland Interconnection, LLC) in the US opened most of its ancillary markets to loads in 2006 [101]. The UK was among the first to deploy DR for frequency response using aggregated large industrial loads with under-frequency load shedding, while in 2007 load already accounted for around 30% of the system’s spinning reserves. ERCOT (Electric Reliability Council of Texas) uses

2,400 MW of its demand for spinning reserve, mainly by industrial users with peak demands of 10 MW and more, equipped by under-frequency relays [104].

The potential of industrial loads in providing ancillary services via DR was analysed in [108]. Several DR products were recognised:

- Regulation: response to random deviations in scheduled net load;
- Flexibility: additional load-following reserve for large unforecast RES ramps;
- Contingency: rapid and immediate response to a loss in supply;
- Energy: shedding or shifting energy consumption over time;
- Capacity: alternative to generation.

Typically requested load size for individual users participating in DR is between 3 and 25 MW, while in the case of aggregators this size can be smaller [101]. It was reported in [108] that the minimum average power demand of an industrial user has to be 0.5 MW in order to be considered as an appropriate “candidate” for DR. The power capacity of 0.5 MW corresponds to about 200 HVAC units, heat pumps or water heaters, or an aggregation of about 3,500 refrigerators [78]. The types of load most suitable for capacity and energy DR products are those involved in manufacturing processes which can be turned ON/OFF for extended period of time without modulation. The types of load that can provide all five aforementioned DR programs are those equipped with control devices, i.e., loads that can be modulated (fans, pumps, air compressors, etc.), and those that participate in continuous processes (furnaces, smelters, electrolytic cells, etc.) [108]. Thermostatically controlled loads (TCLs) have been recognised to have great potential for fast frequency regulation due to their large number and ability to be turned ON/OFF simultaneously [78].

In the UK and Nordic countries DR for ancillary services initially started with large industrial users with SCADA telemetry [101]. However, with the markets being developed around aggregators, the minimum size requirement has decreased. Aggregators such as ENGIE (former Gaz de France), have successfully included other types of smaller load, namely dual-fuel boilers, back-up generators, and combined heat and power facilities to provide operating reserves that can meet all the technical requirements of the system operator. Pilot sites in Nordic countries investigated aggregated heating loads (water and space heaters) which make 16,000 MW for reserve and balancing. 80% of flexible customers in the UK are contracted via an aggregator, while some also contract directly with National Grid (20%) and through their DNO (10%) [109]. According to an Ofgem survey, there is a great untapped flexibility potential in the UK (around 3GW for reducing demand and around 2 GW for increasing demand), with motors and pumps being the most flexible, followed by lighting, although it is not used frequently [109].

Ancillary services which are currently provided to the National Grid via aggregators are as follows [109]:

- Balancing services – firm frequency response (FFR), with response time within seconds;
- Balancing services – reserve (reducing/increasing/shifting consumption), with response within minutes;
- Capacity - reducing/increasing/shifting consumption when electricity demand is higher than available generation. In order to ensure security of electricity supply, payments are provided to existing and prospective generators and demand side providers, in return for a commitment to provide capacity during a system stress event;
- Demand Turn Up – shifting demand to the periods of the day when RES production is higher.

1.5.4 Summary of past work

Previous sections have provided an overview of the state-of-the-art of research and industrial practices in the area of demand observability and DSM motivated by either market or network based applications. Some of the main research problems and points identified in the literature review are summarised as follows:

1) Large-scale DSM is getting more attention as one of the flexibility providers at distribution and transmission network level, which brings the need to characterise aggregate demand at network buses with respect to its flexibility, voltage or frequency dependence (i.e., load response to voltage and frequency deviation) [10, 62, 70, 74-79]. Apart from real-time

estimation, forecasting demand flexibility, e.g., day ahead, is of particular importance for any DR program that is planned in advance [67].

2) Non-intrusive load monitoring (NILM) methods have shown to be useful for demand disaggregation, however the requirement for high granularity of measurements poses an issue for existing communication infrastructure in distribution networks, especially for disaggregation of a large number of users [33, 36-38]. Therefore, applicability of SMs combined with sub-metering technologies as an alternative for aggregate demand disaggregation should be investigated. In addition, SM data compliance with the information and operation needs of the DSO in smart grid environment should be examined.

3) Most DR programs are motivated by either economic benefit, maximising the use of renewables, or deployment of DR in contingency situations [88, 94, 95]. Considering network performance as a constraint in daily planning of DR is usually missing.

4) DSO is seen as one of the main flexibility providers supporting the TSO [43], apart from aggregators [71]. In many DSM programs [70, 84, 85], the aggregate demand is scheduled with the aim to follow a pre-defined loading curve.

5) Load margin in the distribution network is often neglected compared to the transmission network analysis. With the proliferation of DER at the distribution side, the capability of the distribution network in providing services to the transmission network is becoming more critical [43]. New large types of load, such as EVs and heat pumps [15], will bring new challenges with respect to loadability of the distribution network.

6) Optimal Allocation of DGs and DR resources for improving voltage stability has been performed [20, 22, 49], however optimal allocation of load dispatch for voltage stability improvement with respect to static and dynamic load components has not been analysed.

7) In various analyses focusing on DR, voltage stability and load margin, load is often modelled using the constant power model and its voltage sensitivity

is neglected. This approach does not account for the potential changes in network power flows and consequently voltage stability, arising from different load control actions over static and dynamic loads at a network level. It is important to note that not only composite model should be used, including static and dynamic load components (as recommended in [43, 44]), but also the daily and seasonal changes in load model parameters should be observed.

1.6 Aims and objectives

Based on the identified research gaps of the research in past work, two main research questions have been extracted which will be addressed in this thesis:

Research question 1: *How can we use the existing smart metering technologies to estimate/predict aggregate demand flexibility and derive time-changing shares of different load components?*

Research question 2: *Can a DR program be "tailored" to meet the load profile requirements of the transmission or distribution network operator, available demand flexibility, while maintaining/improving loadability of the distribution network?*

1.6.1 Aims of the research

There are two main aims of the research presented in this thesis, addressing the identified research questions. The first one is to use data mining to forecast (day-ahead) or estimate (in close to real time) the composition of aggregated demand in residential sector with limited demand observability enabled by SMs. This methodology is referred to as Advanced Demand Profiling. The second aim concerns optimised load scheduling in distribution network as a support to the transmission network, taking into account the requirements of the network operator, the limitations in load flexibility and preservation of network performance indicators, in this case load margin of the distribution network. This methodology is referred to as Multi-objective DSM.

1.6.2 The overview of the research

Advanced Demand Profiling methodology enables decomposition of forecast real and reactive demand at the aggregation level (e.g., a substation) using artificial neural networks (ANNs). The methodology builds on historical SM data and appliance level (sub-metering) data from a limited number of residential end-users, and results in load shares of six pre-defined load categories within the total aggregated demand. In addition to limited observability, the impediments taken into account are missing data and different sampling steps of different SMs. The data are therefore pre-processed before aggregation in an off-line manner, although future development could include methods for on-line processing and restoring on-line data streams with missing values. The analysis of results provides an important piece of information – it defines the minimum percentage of end-users that has to be observed by SMs having sub-metering functionality in order to allow for a confident prediction of the demand composition of the overall demand (including both observed and non-observed users).

Multi-objective DSM feeds on the output of the first methodology, i.e., the time-varying demand composition at each load bus is used to model the load at each load bus using composite (ZIP+IM) load model, where the shares of the load model components change following the changes in demand composition during the day. Information about demand composition is used not only to plan a DR action and predict the behaviour of demand during the load payback at different load buses of the network, but also to assess the network performance indicators (in this case, load margin) affected by the changes in load flows coming from the changes in the size and composition of demand. The controllable components of the composite load model can therefore be optimally scheduled (disconnected at one time step and reconnected at another time step of the planning horizon) to preserve or improve the network performance after the DR action.

1.6.3 The scope of the research

The methodology for demand decomposition, which is developed in the first part of the research, provides information about the shares (in percentage or per unit) of different load categories (e.g., induction motors, resistive loads, etc.), but not of individual appliances (e.g., washing machines, water heaters, etc.). Load appliances belonging to the same category have similar steady-state and dynamic voltage-dependent load characteristic. Rather than disaggregating individual user's daily demand into electrical appliances, as in [36, 37, 110], the methodology aims at decomposing the aggregated demand into load categories and controllable load, similarly to [39, 42, 86]. This type of information is deemed acceptable for the DSO (or other DR responsible party), as it classifies load flexibility into load categories with similar static and dynamic behaviour. As a consequence, aggregated demand flexibility is seen as a type of DER. The ultimate aim is to establish what percentage of users would have to be monitored using SMs with sub-metering technologies in order to provide relevant information to the network operator for efficient deployment of a DSM program.

After a DR action is triggered, shifting of load will change load composition, and thereby potentially the steady state or dynamic response of aggregated demand in case of a small or large network disturbance (e.g., voltage step change due to transformer tap changes, system faults, etc.), which might affect the angular and voltage stability of the power system [39] – this effect would be emphasized at transmission system level. Information about the load composition can be highly useful in these cases, as it can be used for: *i)* estimation of the dynamic load response at some given time; *ii)* prediction of the load response at some point in the future based on readily available information without having to perform field tests or measurements [16]. The aim is to ensure desired (or maintain existing) dynamic response of demand at given hour following the shift of the demand.

Any wide-scale DR program, which changes the load profile across the whole network or network area, should be complemented by appropriate network performance analysis. The type of the analysis will depend on the type of the network and its voltage level. The network performance indicator, chosen for illustration purposes in this thesis, is load margin of the distribution network. However, other indicators, such as different aspects of angular or frequency stability, or network losses, could be used individually or in combination, to accomplish a network performance-aided DSM. The methodology for multi-objective DSM presented in the second part of the thesis can be easily transferred to transmission network with inclusion of other indicators of network stability and security. Finally, the methodology is seen as a decision support tool for a network operator or DR responsible entity, when planning (in short term) a wide-scale DSM program.

This work does not analyse ways to incentivise end-users to participate in DSM. It does, however, assume a direct load control program for load shifting that would ensure a more confident response from the demand-side. The final outcome of a DR program depends not only on the flexibility of demand, but also on its availability, i.e., on willingness of the end-users to participate in DR. Therefore, different levels of participation of the end-users will be taken into account during the analysis of multi-objective DSM.

1.6.4 Objectives of the research

The objectives of the research contributing to the aforementioned aims are given as follows:

1) To explain the reasons for the use of big data analytics in distribution network studies and give an overview of data mining methods typically used in power network studies;

2) To examine the possibilities offered by the availability of SM data with respect to present and future data requirements of the DNO;

3) To illustrate application of data mining methods for knowledge extraction from example database of a real distribution network;

4) To develop a methodology for aggregated demand decomposition using conventional SMs and a limited number of SMs having sub-metering functionality;

5) To develop an effective way to aggregate SM data streams containing missing samples and arriving to the data concentrator point at different sampling steps (one, ten, thirty and sixty minutes);

6) To illustrate the methodology on a realistic dataset comprising a large number of aggregated residential end-users, using a UK statistics-driven per-appliance consumption model adopted from [111];

7) To develop a methodology for estimating reactive load data, when these are not available from existing SMs, based on real power measurements and sub-metering data;

8) To test the accuracy of the methodology on a number of cases with different SM coverage, i.e., with different portion of end-users monitored with SMs

having sub-metering functionality. The purpose of the analysis is to investigate the minimum required SM coverage (in %) of end-users connected to the same bulk point in a residential area necessary to obtain the desired accuracy of both real and reactive demand decomposition of the overall aggregation;

9) To validate the methodology and achieved confidence level using a dataset from a real pilot site with aggregation of end-users with SMs and sub-metering data;

10) To compare the developed methodology to a time series method and evaluate the effect of additional data types in the input data on the accuracy of the methodology;

11) To develop a graphical user interface (GUI) for illustration of demand decomposition results foreseen as a decision support tool for DR responsible party;

12) To use demand decomposition results to model demand at distribution network load buses using appropriate realistic load model that accounts for the change in demand composition during the day;

13) To develop a methodology and a corresponding optimisation method for optimal scheduling of controllable demand in distribution network with three objective functions: *i)* meeting the pre-defined loading curve, *ii)* maintaining or improving load margin of the distribution network and *iii)* preserving demand composition of aggregated demand;

14) To illustrate the developed methodology on a range of case studies using representative distribution network model.

1.7 Main contributions of the research

The main contributions of the research presented in this thesis are in the area of demand profiling (disaggregation) and optimal demand side management. The following points summarise the main contributions of the research with the numbers given in parentheses corresponding to the relevant research publications by the author where these results are presented (the full list of thesis based publications is given in the Appendix B):

1) A critical overview of different data analytics methods, including pioneering use/discussion on use of text mining in power system studies, and their possible applications in distribution system studies is provided, followed by an illustrative example of data mining application for distribution network asset management. {B3, B6, B7}

2) A comprehensive overview of data requirements for present and future power network studies and network operation is given, with a special focus on distribution network. In this context, an analysis of smart metering technology and its advantages and disadvantages with respect to improved observability of the distribution network is also provided. {B4}

3) A probabilistic methodology is developed for derivation of reactive demand data for an aggregation of users based on available real power measurements by SMs, as in many cases SMs measure only real power. This solution facilitates not only better observability of the distribution network with respect to reactive load flows, and consequently, more accurate load modelling, but also an assessment of the power factor of demand. {B1}

4) A pioneering methodology is developed for time varying decomposition of forecast real *and* reactive power demand, using SMs with sub-metering functionality and pre-trained ANN. The decomposition is performed into six load categories and controllable/uncontrollable load in residential district. As a part of this methodology, an effective way for data restoration of missing samples prior to aggregation of SM data streams is also provided. Considering all the obstacles in deploying SMs, and in particular those with sub-metering technologies (whether they are intrusive or non-intrusive), the research identifies the minimum share of demand that needs to be monitored to “per-appliance level” in order to obtain confident information about composition of the load, and more importantly load controllability in the area, so that efficient DSM programmes can be applied. {B1, B5, B8, B9, B10, B12}

5) A GUI is developed for visualisation of the results of forecast total and decomposed real and reactive demand in a distribution network control

centre, foreseen as a practical decision support tool for short-term DR planning. {B10}

6) Particle swarm optimisation – based load scheduling methodology is developed for meeting multiple objectives of the DNO – meeting the target loading at the GSP, keeping/improving loadability of the distribution network, and maintaining the composition of demand at the GSP. Relying on a more accurate load modelling provided by the methodology for advanced demand profiling, the DSM program can be planned by taking into account the forecast flexibility of the aggregate demand at each load bus. {B2, B11, B13, B14}

7) Unlike previous work on DSM, the proposed methodology schedules, optimally and simultaneously, *two* controllable load types, namely constant impedance load and induction motors, so that, in addition to meeting target loading at GSP, the load margin after the DSM program is at least maintained, if not improved, or that load-follows-generation approach is facilitated, or that other aspects of distribution or transmission network performance are maintained or improved. {B2, B11, B13}

1.8 Thesis overview

The thesis is organised into five chapters. This chapter (*Chapter 1*) is the introductory chapter, while the overview of the remaining chapters is given below.

Chapter 2 The Need for and Application of Data Analytics in Distribution System Studies

This chapter explains the importance of data mining methods in power system studies, mainly focusing on the distribution network. It provides an overview of the smart meter technology, as one of the main enablers of smart grid evolution and activation of the demand side in network daily operation. Present and future data requirements of the distribution network operator are critically compared against the data types available (now or in the future) from network monitoring systems. A case study illustrating application and benefits of data mining methods for asset management is given at the end of the chapter. In addition, a pioneering discussion on text mining applications in power system studies is provided.

Chapter 3 Advanced Demand Profiling

The third chapter of the thesis introduces the methodology for decomposition of aggregated residential demand using smart meter data and artificial neural networks. It presents several steps in the methodology, from data aggregation and pre-processing, to the application of artificial neural networks and decomposition of active and reactive load at the aggregation level. An approach for obtaining probabilistic aggregated reactive load curve is also discussed as a solution to the lack of reactive load measurements at the end-users' point. A case study validating the approach on an actual dataset from a pilot site is presented. Finally, a graphical user interface for advanced demand profiling is introduced, as a practical tool foreseen to be used in a distribution network control centre for short-term DR planning.

Chapter 4 Multi-objective Demand Side Management at Distribution Network Level

The chapter presents a comprehensive methodology for optimal scheduling of distribution network loads in support of transmission network operation. The objective of the proposed DSM program is the load profile shaping, as a balancing service to be offered to the TSO while maintaining the load composition and one or more network performance indicators to values they had prior to the DSM action. The case study uses the IEEE 33 bus distribution network and takes into account influence of load modelling, limited demand flexibility and load payback, illustrating the importance of considering realistic assumptions when estimating the success of a DSM program.

Chapter 5 Conclusions and Further Work

The last chapter of the thesis provides major conclusions of the research, and suggestions for further work and development in the area of load profiling and DSM.

2 The Need for and Application of Data Analytics in Distribution System Studies

2.1 Introduction

“Although the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant” [112]. With the introduction of information and communication technologies (ICT) and significant deployment of monitoring systems resulting in large amount of data streams, the need for utilization of data mining techniques has increased. Even though the data volumes power industry is dealing with do not compare to those used by Internet, for example, the number of data sources in transmission and distribution systems is continuously growing and filling the databases of power utilities with data that are much bigger than it used to be the case. Although there is an obvious need for increasing the size of the existing databases to be able to accommodate new static (e.g., reports) and dynamic data (e.g., real-time measurements), an important question is if and to what extent the already existing data is being harnessed. In other words, how useful is the data that is already being collected, and can some data mining methods facilitate the usefulness, i.e., the knowledge extraction from the existing and new types of data? This chapter therefore analyses the present and future data needs for the distribution system studies.

Furthermore, the chapter revises commonly used data mining methods in power system studies, with an emphasis on their application in distribution network studies. Following this, the benefits, both present and potential, coming from smart meter (SM) rollout are investigated. An important benefit should be brought to energy suppliers, who will have access to remote monitoring of the end-users, and improved bidirectional information exchange with the users. At the same time, the end-users should be able to save energy and benefit from reduced energy bills. Therefore, SMs will facilitate development of smart(er) grids and contribute to the development of low carbon policies.

Finally, a case study is presented, illustrating how preventive maintenance of distribution utility feeders can be facilitated by extracting useful information (knowledge) from raw historical data about faults on feeders in HV and LV distribution network.

2.2 Data analytics

Big data is defined as data that are high in either volume, variety, velocity, veracity, or all four of these features [30, 113]. These four features are known as the *four Vs* of big data. Even though the data volume power utilities are dealing with may not be as large as in other domains, such as Internet, it is constantly growing and needs to be efficiently handled and processed in order to be useful (the *fifth V* can be defined as the value of big data analysis [30]). Variety refers to different types of data (textual, numerical, etc.), while velocity refers to the speed at which data is coming and at which it needs to be processed. Veracity is reflected in uncertain or missing data. Data mining techniques have already been widely used in power industry for power system security assessment, fault detection, control, load and price forecasting and power generation risk management [114]. It was reported in [115] that the areas with highest priority for data analytics applications in power engineering are energy forecasting, SM data analytics, asset management, network operation and customer segmentation. The two basic tasks for data mining methods are prediction, based on observations of already existing records, and knowledge discovery from big databases [116]. The main challenges arising with large databases are the following [30, 117]:

- Database volume, which is equal to the product of the number of instances (recordings or measurements) and dimensionality of data (attributes) describing those instances;
- Speed of data acquisition and update;
- Variety of data sources and data formats;

- Incompleteness of data;
- Quality and security of data;
- Benefits or usefulness obtained from the data analysis.

Each instance (measurement or recording) is characterized by the values of attributes that measure different aspects of the instance. There are several types of attributes, although typical data mining methods deal mostly with numeric and nominal, or categorical ones [118]. Additional types of data that may be of interest in power utilities are images (e.g., thermal camera recordings of power system assets), textual data (from different reports that may be off-line or on-line) or voice recordings (from customer services, for example) [113].

One of the challenges of big data analytics is to make correlations between different databases, e.g., between weather and network outages, and use this knowledge to prevent further faults and disruptions in the system. Different departments in distribution utilities use different styles of record keeping, conventions, time periods, levels of data aggregation, and different identifiers, and will have different types of errors. All these are aggravating factors for connection and correlation of databases. This problem is also referred to as entity identification problem [119]. The data has to be assembled, integrated, and cleaned up, taking into account the importance of the right type and level of aggregation of data, prior to any future processing [118, 120].

Any knowledge derived from databases should bring novelty and also be valid, useful and presented in a simple way [121]. The number of data mining methods and their modifications depending on the application has been constantly increasing, which can bring confusion in setting clear boundaries between them. Nevertheless, three groups of data mining methods are considered essential: correlation, regression and classification [114, 122]. These will be discussed in more detail in the following sections.

2.2.1 Correlation

Correlation is a widely used statistical tool for retrieving relationships between data. In the case of linear correlation, it gives the strength and direction (positive or negative) of the relationship between numerical variables. Also, as a means of feature (attribute) selection, it can be very useful for rejecting uncorrelated data (or, on the contrary, highly correlated data, where in case of two variables, one can be rejected as redundant

[120]), i.e., reducing data size. This makes it a very important step in the data pre-processing, i.e., cleaning the data for future classification. The typical measure of correlation is Pearson's coefficient r [114], calculated as follows:

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sqrt{[\sum_{i=1}^{n}(x_i-\bar{X})^2][\sum_{i=1}^{n}(y_i-\bar{Y})^2]}}, \ -1 \leq r_{X,Y} \leq 1 \tag{2.1}$$

where $cov(X,Y)$ is the covariance of variables X and Y, σ_X and σ_Y are their standard deviations, and $\bar{X}$ and $\bar{Y}$ are the mean values of variables X and Y, respectively. If the Pearson's coefficient is equal to zero, it means that the two variables are independent, i.e., there is no correlation. The closer the coefficient is to unity, the stronger the correlation is, with the sign defining the direction of correlation. This measure can only represent the linear correlation, and is not robust to outliers [114]. If the mutual relationship is nonlinear, the Pearson's coefficient is not appropriate for description of the strength and direction of the relationship. Hence nonlinear regression is used to find the relationship between variables in cases like this.

In case of nominal (categorical) attributes, it might be convenient to use chi-square χ^2 (the Pearson statistic) [119] as a measure of correlation test. It can be used to investigate the correlation between, e.g., type of a feeder and class of customers connected to it. If there is a higher correlation between certain feeder type and a customer class, it means that by the customer class one can assess, with higher probability, the type of a feeder it is connected to, and vice versa. Let there be two attributes A and B, e.g., type of a feeder and customer class, with c and r being the number of possible categories of the attributes, respectively. If possible values for A are $a_1, a_2, \ldots, a_c$, and possible values for B are $b_1, b_2, \ldots, b_r$, chi-square is calculated as follows:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \tag{2.2}$$

where o_{ij} represents the actual frequency of the joint event (A_i, B_j) where A takes the value a_i and B takes value b_j, while e_{ij} shows the expected frequency of the joint event, calculated as:

$$e_{ij} = \frac{count(A=a_i)\times count(B=b_j)}{n} \tag{2.3}$$

with n being the number of data pairs (a_i, b_j) in dataset and $count(A = a_i)$ shows the number of events when $A = a_i$, similarly to $count(B = b_j)$. The Pearson statistic tests the hypothesis that the attributes A and B are mutually independent.

2.2.2 Regression

Linear regression is a well-known technique for numeric prediction, used for finding numerical relations between a numerical response attribute (e.g., number of faults per feeder) and numerical predictor attributes (e.g., dimensions of a feeder, number of customers connected to it, etc.). The response (y) is presented as a linear combination of predictors $(x_1, x_2, \dots, x_n)$ and weights or regression coefficients $(w_0, w_1, \dots, w_n)$, given in the following form [119]:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{2.4}$$

Weights are calculated based on the training data, i.e. a given set of examples of response values and corresponding predictors' values. Once calculated, these numerical weights can be used as predictors of the unknown outcome if the predictor attributes are known [118]. In this sense, linear regression can also be used for data cleaning, i.e., filling of missing values in a dataset, as an important part of data preparation for further analysis. In cases of data with nonlinear dependency, where linear regression gives only a rough estimation of the prediction function, more accurate estimation is made using non-linear regression model given in the following form [123]:

$$y_i = f(x_i, \theta) + \varepsilon_i \ , \tag{2.5}$$

where y_i and x_i are vectors of response and predictor attributes in the i-th instance, respectively, θ is the vector of weights, while ε_i is a random error. The weight vector that is unknown can be estimated from the training set using least squares method, i.e., minimization of the following expression:

$$\sum_{i=1}^{n} (y_i - f(x_i, \theta))^2 \ . \tag{2.6}$$

If the response variable is nominal (categorical), it is preferable to use the multinomial logit model, which gives the relative risk of being in one category against being in the reference category k expressed as a linear function of predictor variables [122]. The probability of each outcome is given as a nonlinear function of p predictor variables (attributes). The model is given as follows:

$$\begin{cases} \ln\left(\frac{\pi_1}{\pi_k}\right) = \alpha_1 + \beta_{11}X_1 + \cdots + \beta_{1p}X_p \\ \vdots \\ \ln\left(\frac{\pi_{k-1}}{\pi_k}\right) = \alpha_{(k-1)} + \beta_{(k-1)1}X_1 + \cdots + \beta_{(k-1)p}X_p \end{cases} \tag{2.7}$$

where $\pi_j = P(y = j)$ is the probability of the response variable y being in category j and k is the number of response categories. β coefficients are estimated by solving the system of $k - 1$ equations. Coefficient β_{ji} expresses the fact that probability of the response variable being in category j compared to the probability of being in category k increases exp(β_{jk}) times for each unit increase in X_i, having all other predictor variables constant.

2.2.3 Classification

Classification is a general term for all data mining methods that form groups (classes) of data based on some categorical rules [119]. It is a two-step process: first, in the training step, a model, i.e., a number of classes with defined attributes is formed based on available observations (patterns, data items or feature vectors). The second step is to classify unseen examples based on their attributes. In supervised classification methods, a given set of labelled patterns (training data) is used to learn description of each class (group). In other words, grouping of new data is supervised by the training data [114]. Thus, description of a new pattern associates it to one of the predefined classes [124]. The aim of supervised methods is to build a model that makes predictions based on evidence in the presence of uncertainty [122].

Classification usually analyses data given in a vector form, having either continuous or discrete values [114]. The first part of data classification, as in any other data mining process, is pre-processing of raw data, which consists of several stages [121]:

- Extraction of the data that can be useful for further analysis;
- Removing data noise;
- Statistical analysis for generating new useful variables;
- Organising data in a form suitable for the desired classification method.

Some of the commonly used classification methods applied in power system studies are decision trees, artificial neural networks (ANN) and Bayes classifiers.

2.2.3.1 Decision trees

This method is based on generating comprehensive rules for dealing with both continuous and discrete data. The tree structure consists of if-then rules, i.e., tests on attributes given in nodes, branches that represent results of the test classes, and leaves containing class labels [125]. Class labels can be nominal, in case of classification trees, or numerical, in case of regression trees [122]. The ending leaf class value of a regression tree refers to the average value of all the instances reaching that leaf [118].

An important step in constructing a decision tree is deciding on the attribute that will be tested in a node (i.e., splitting attribute) and defining further partition of the set of instances into subsets (classes). Methods for this (information gain, gain ratio and Gini index) are described in more detail in [119]. Depending on the type of splitting attribute, the test is given in the following form:

- If the attribute is given as a discrete value, then possible output branches correspond to all values of the attribute;
- If the attribute is a continuous value, then there are two branches referring to values under or above a certain splitting point (usually taken as a midpoint of two known adjacent values of the attribute);
- In case of a binary tree, where the attribute is a discrete value, the test is formed based on the condition $A \in S_A$, where S_A is the splitting subset for the attribute A. Possible output branches correspond to answers *yes* (attribute belongs to the subset) or *no*.

A very important advantage of decision trees is their capability of handling multidimensional data [119]. The computational complexity is given as $n \times |D| \times log|D|$, where n is the number of attributes and $|D|$ is the number of training instances in the dataset D [119].

2.2.3.2 Artificial neural networks

Artificial neural networks (ANNs) present an upgrade of logistic (nonlinear) regression [118]. ANNs are useful in cases where [126]:

- Examples of predictive and response variables exist, but their relationship cannot be derived using algorithmic solutions;
- The relationship changes over time, i.e., the solution has to be adapted to the change.

Although there are various types of ANN, they are all based on the single-layer perceptron, depicted in Figure 2.1 [127]. A perceptron is a binary classifier, i.e., a function deciding whether an input belongs to a class or not. It is based on a nonlinear model of a brain neuron. For a classification into more than two classes the number of neurons in the perceptron expands. The summing node of the perceptron computes a linear combination of the inputs (denoted as $x_1, x_2, \dots x_m$) using a set of weights (denoted as $w_1, w_2, \dots w_m$) and assigns a bias (fixed input), denoted as b, to it. The resulting output is then forwarded to a transfer function (hard limiter in Figure 2.1, denoted as φ). The input v to the transfer function is defined as:

$$v = \sum_{i=1}^{m} w_i x_i + b \tag{2.8}$$

In the simplest case of binary classification, the result of the transfer function is +1 if the output v (i.e., transfer function input) is positive and -1 if it is negative.

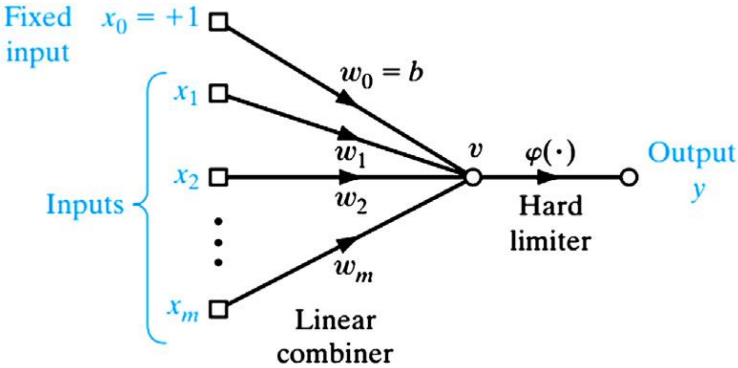


Figure 2.1 Signal flow of the perceptron (adopted from [127])

The perceptron’s task is to classify the set of inputs into class 1 if the result of the transfer function is +1 and class 2 if the result is -1. If the classification is not binary, a differentiable transfer (or activation) function is used, limiting the amplitude of the output of the neuron (v). Sigmoid transfer functions are commonly used for pattern recognition, while linear functions are used for function fitting [128]. Sigmoid function can be given in the following logistic form [127] :

$$\varphi(v) = \frac{1}{1+e^{-av}} \tag{2.9}$$

where a is the slope parameter of the sigmoid function, as depicted in Figure 2.2.

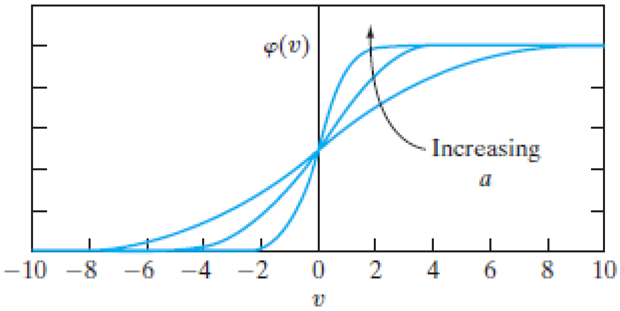


Figure 2.2 Sigmoid function (adopted from [127])

The weights of the perceptron, which are initially unknown, are adjusted iteratively, during the training process. The training process involves determining an optimal set of weights based on the observed examples (inputs and corresponding known outputs - targets).

The most common type of ANN is the multi-layer perceptron, consisting of an input layer, hidden layers and one output layer [129]. Each layer consists of nodes, whose number in the input layer corresponds to the number of inputs, while in the hidden and output layer it corresponds to the number of hidden and output neurons, respectively. The hidden neurons, which are not "visible" from the input and output layers, perform a nonlinear transformation of the input signals, which may characterise the training data in a way that was not obvious in the original input [127]. This characterisation evolves through the training process of the ANN.

Let us observe a simple two layer feed-forward (information flow is from input to output layer only) ANN with backpropagation. The two layers refer to the fact that the ANN contains one hidden and one output layer (the input layer does not count as it is not performing any computation). In multi-layer ANN each layer has as the input the output of the preceding layer. During the training process the weights, whose values are initially randomly generated and assigned to the input, are adjusted iteratively by comparing the resulting output with known target values and returning the error backwards, through the hidden layer, to adjust the weights values. Once the error is smaller than the given threshold, or the predefined maximum number of iterations is reached, the training stops. After being trained, the performance of the ANN can be validated either with the same set of input data (where the ANN output is compared with the target) or with a new test data set containing known inputs and outputs (where the ANN output is compared to the known test output).

The main disadvantages of ANNs are the empirical design of network structures and parameters, over-fitting and the need for numerous training instances [114, 130]. Apart from feedforward networks, there are also recurrent networks, which have at least one feedback loop – for example, each neuron in a layer can feed its output to the input of all the other neurons [127]. Self-Organising maps (SOM) are another kind of neural network that performs clustering analysis of the input data. The basic units are the neurons organised in two dimensional layers: the input layer, and the output layer, which is often referred to as the output map. All the input neurons are connected to all the output neurons, and these connections have "strengths" associated with them [28]. The output neuron with the strongest response is said to be the winner, and is the answer for that input.

2.2.3.3 Bayes classifiers

Another way to classify data is using Bayes classifiers that have shown good management of datasets with large number of predictors and the advantage of requiring a small amount of training data to estimate the parameters necessary for classification [122]. Naïve Bayes classifiers assume that the attributes describing an instance (feature) are mutually independent – this assumption is called "class-conditional independence" [119]. The Bayes theorem is given as follows [119]:

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)} \tag{2.10}$$

where $P(C_k|X)$ is "a posteriori" probability of instance X belonging to the class C_k of the total of K classes. $P(C_k)$ is the "a priori" probability of the class C_k , i.e., probability that an instance belongs to the class C_k, regardless of its attributes values, $P(X)$ is "a priori" probability of instance X in the dataset, and $P(X|C_k)$ is "a posteriori" probability of the instance X having as a condition existence of class C_k.

Bayes classifier assigns instance to a class with the highest "a posteriori" probability, i.e., it maximises $P(C_k|X)$, where $k = 1, \dots, K.$ An instance X belongs to a class C_k if and only if $P(C_k|X) > P(C_j|X)$ for $1 \leq j \leq K, j \neq k.$

Since $P(X)$ is constant regardless of the class, what should be classified is the numerator $P(C_k)P(X|C_k)$. If prior probabilities of classes are unknown, they are presumed to be either equal or estimated by $P(C_k) = |C_{k,D}| \,/\, |D|$, where $|C_{k,D}|$ is the number of training instances of class C_k in training set D [119]. Given the assumption that all the attributes values in an instance are mutually independent, it follows that:

$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \tag{2.11}$$

Probabilities of the attributes conditioned by the class C_k can be estimated based on the training set. Estimation is done according to the type of attribute:

a) If the attribute is nominal (categorical), $P(x_i|C_k)$ is the number of instances in the training set assigned to class C_k, having the value x_i for the attribute A_i, divided by the number of instances of class C_k in the training set D ($|C_{k,D}|$).

b) If the attribute has a continuous numerical value, it is considered to have a Gaussian distribution with a mean μ and standard deviation σ, defined as:

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.12}$$

So, probability of an attribute value conditioned by a class C_k is given as follows:

$$P(x_i|\, C_k) = g\left(x_i, \mu_{C_k}, \sigma_{C_k}\right), \tag{2.13}$$

where μ_{C_k} and σ_{C_k} are mean value and standard deviation, respectively, of attribute A_i for training instances of class C_k.

Theoretically, Bayes classifiers have the minimum error rate compared to other classifiers [119].

2.2.4 Clustering

Clustering is a common name for the group of unsupervised data mining methods, which can also be seen as a subgroup of classification methods. It groups patterns (observations, data items, or feature vectors) into groups (clusters) with an aim of hypothesis formation or decision-making. Classification of measurements is based on either (*i*) goodness-of-fit to a postulated model, or (*ii*) natural groupings (clusters) revealed through analysis. While classification models assign new data to previously defined classes which are specified as a target, clustering models do not use a target.

Clustering, as an unsupervised method, distinguishes clusters dynamically – in other words, "category labels are data driven" [124]. The process of clustering is very subjective, which means that the same set of data can be partitioned (clustered) differently using different methods. That is why one has to be very careful with the choice of approach. An essential advantage of clustering is that it can extract groupings automatically - "learning by observation rather than learning by examples" [119].

When choosing the optimal clustering approach, two criteria are followed [28]:

- Compactness: members of each cluster should be as close to each other as possible;
- Separation: the clusters should be widely spaced from each other.

Clustering is useful for data exploration – if there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. It can also serve as a useful data pre-processing step to identify homogeneous groups on which to build supervised models. Clustering may be used for identifying anomalies - once the data has been segmented into clusters, cases that do not fit well into any clusters are considered as anomalies or outliers [131].

There are several different clustering approaches depending on the definition of the data and on the clustering mechanism [30]:

- Hierarchical clustering, which further can be divided into:
    - Agglomerative ("bottom up") approach, where the N patterns (samples) are initially considered as single clusters, and pairs of clusters are merged until the required final number of clusters is reached. This requires high computational complexity, given with $O(N^3)$;
    - Divisive ("top down") approach, in which all patterns are included initially into one cluster, and the clusters are progressively split. Computational burden is also high in this case - $O(2^N)$ or $O(N^2)$ in special cases;

- Partitional (centroid-based) clustering: each cluster is represented by a centroid, which may or may not belong to the dataset. An example is k-means clustering, whose computational complexity is given with $O(N)$ [124];
- Distribution-based clustering: clusters are formed from data having the same distribution. Gaussian mixture models belong to this group;
- Density-based clustering: clusters are formed of data areas with higher density than the rest of the dataset. This method is primarily used for finding clusters of non-spherical shape [30];
- Information theory-based clustering: clusters are formed on the basis of the modes of the probability density function of the initial dataset.

Hierarchical clustering can also be categorised into three groups [119]:

- Algorithmic – considering data objects as deterministic and computing clusters according to the deterministic distances between objects;
- Probabilistic – in case of missing data, this method uses probabilistic methods to measure distance between objects;
- Bayesian – advanced method for presenting distribution of possible clusterings, i.e., a group of clustering compositions and their probabilities.

Similarity between data is commonly evaluated by distances, where distance d between two vectors (patterns) x_j and x_k is usually measured through the Euclidean distance [25]:

$$d(x_j, x_k) = \|x_j - x_k\|. \tag{2.14}$$

Some other similarity measures for objects described with numerical attributes are Manhattan and Minkowski distances [119]. Euclidean distance is the most common measure of similarity for patterns of continuous variables [132]. For the non-numerical data, a convenient method for calculating the distance is Hamming error, detailed in [133].

In the bottom-up approach, vectors with the smallest distances are merged together into one cluster. In the next step, distances between the newly formed clusters and all other vectors are computed resulting in new clusters based on the smallest distance. The process continues repeatedly until only one cluster remains, visually presented as a cluster tree (Figure 2.3 [29]). Since formed clusters could be merged into one based on their similarity, it is easy to overestimate the optimal number of clusters by this method [25]. The last step is division of the cluster tree into coherent groups based on a similarity criteria (vertical axis in Figure 2.3), usually the distance between clusters or inconsistency measure [25]. Final number of clusters depends on this threshold – the higher it is the smaller is the number of clusters.

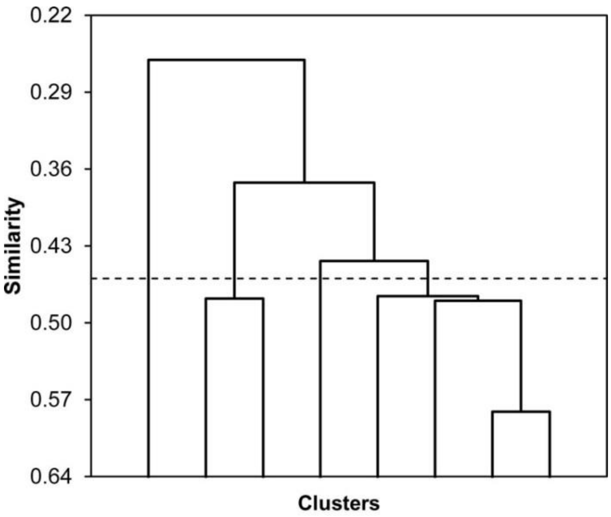


Figure 2.3 Example of a cluster tree (dendrogram) in hierarchical clustering (adapted from [134])

Partitional methods have the advantage in applications involving large datasets for which the construction of a dendrogram is computationally too demanding. On the other hand, the main drawback of all partitional methods is the need for defining the desired number of clusters in advance [124]. Optimal number of clusters or the quality of clustering can be tested in two ways: by extrinsic means (i.e., comparing clusters with a ground truth, e.g., comparing customer grouping with already existing tariff profiles) or by intrinsic means (i.e., examining the compactness of individual clusters and separation between clusters) [119].

Apart from the basic methods described above, there are various modifications in clustering, including: “follow the leader” procedure (FDL), Self-Organising Maps (SOM), Gaussian Mixture Model (GMM), k-means (KM), Fuzzy C-means (FCM), Support Vector Clustering (SVC), Ant Colony Clustering (ACC), Renyi Entropy-based Clustering (REC), etc. k-means and different variations of ANN have been most widely used clustering methods in big data analytics [124].

2.2.4.1 k-means clustering

k-means is a classical clustering method [118], where initial centres of k clusters are randomly chosen from the data set. All other data objects (instances) are assigned to their closest cluster centre according to the ordinary Euclidean distance metric ($\|x_j - x_k\|$, in case of two vectors or patterns x_j and x_k). In the next step the centroid, or mean, of the instances in each cluster is calculated [28]. These centroids are taken to be the new centre values for their respective clusters. The whole process is iteratively repeated with the new cluster centres, until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centres have stabilized and do not change any more. The main drawback of this method, as most of the clustering methods, is the need for predefining the expected number of clusters.

2.2.5 Comparison of data analytics methods

In order to further analyse performance of data mining methods applied specifically to big data systems, comparison of the previously discussed methods is given in Figure 2.4, in a form of a “radar” diagram. Good performance of the method is marked as 3 and bad performance is marked as 1, based on the comparative analysis given in Table 2.1. Mark 2 was given if the quality of performance was not strictly defined. The following criteria was used when forming the diagram:

1) Influence of the initialization process and the need for training data;
2) Variety of types of data (numerical, textual, nominal/categorical...) that the method covers;
3) How applicable the method is to a very large size data sets;
4) Speed (computational complexity), i.e., memory usage.

As seen from Figure 2.4, decision trees, ANN and Bayes classifiers have the best performance in handling big and heterogeneous data, even though most methods handle big data sets well. k-means also showed the highest computational speed, which justifies its frequent use in big data analytics.

Table 2.1 Comparison of data mining methods

Method	Advantages	Disadvantages

ANN	Deals with large data sets;
	Estimates non-linear relationships;
	Adaptable to changes (new data) in dataset;
	Tolerant to noise [36]
		Requires numerous samples [130];
		Requires retraining

Decision trees	Deals with heterogeneous data; Deals with multidimensional data;	Sensitive to outliers; Computational complexity
Linear regression	Deals with large data sets; Simple to interpret	Discovers only linear relation between numerical variables; Sensitive to outliers
k-means clustering	Deals with large data sets; Low computational complexity	Sensitive to initial cluster centres; Deals only with numerical data; Sensitive to noise and outliers [135]
Hierarchical clustering	No need for predefining the number of clusters	High computational complexity; Cannot deal with large data sets; Prone to overestimating number of clusters
Bayes classifiers	Managing big data sets with high accuracy Requires only small set of training data Minimal error	Sensitive to a-priori probabilities

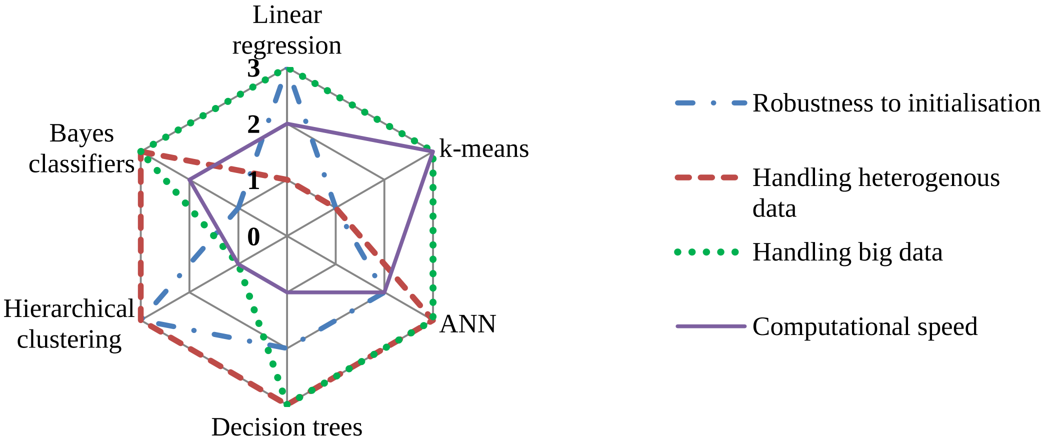


Figure 2.4 Performance comparison of different data mining methods

2.2.6 Text data mining

One of the still undiscovered, or at least, mostly underutilised, possibilities for data mining in power systems analysis is using textual data for knowledge retrieval. This data is usually given in tables or in the form of plain text in reports and articles. Some authors [136] have also pointed out the possibility of using Internet as a large database from which novel information can be extracted. Similarly to other data mining techniques, text data mining includes pre-processing methods, such as text categorization, text clustering and information extracting, as well as data analysis, such as association rules and link analysis [112].

Even though text expresses a vast range of information, it is still very hard to automatically process it because of the form it encodes that information [137]. In the text mining process, textual data is usually represented in a form of feature vectors, where values of features in vectors give a measure of the number of occurrences of a certain term in a document [138]. An example was given in [124] where a number of documents contained several thousand words. Since there was a correlation among words (after eliminating common words such as “the”, “an”, etc.), clusters were formed by groups of words used in a consistent way that happen with a similar frequency in each document. Some authors [138] propose the use of graph models for these purposes, since they can reflect relations between data, e.g., co-occurrence of two or more terms in a document. In a graph model, *n* most frequently occurring terms are given as nodes with their mutual relations presented as edges of the graph. Figure 2.5 [138] gives an example of a graph model.

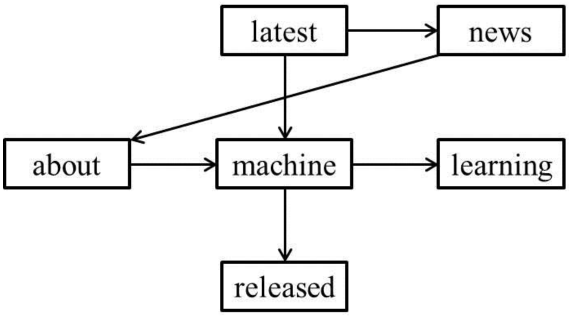


Figure 2.5 Graph model of a text document (adapted from [138])

Many methods have been used for text classification: decision trees, ANNs, nearest neighbour methods, Rocchio’s method, support vector machines, linear least squares, naive Bayes, rule-based methods, etc. [139]. Since there are no tools developed specifically for power engineering studies, an attempt has been made to take advantage of the existing tools for specialized web browsing and data extraction. Two methodologies are explored: one for extracting on-line documents (e.g., journal papers and technical reports from Internet) that are highly related to a specific topic, and the second one for extracting highly related sentences, providing a literature summary on the topic. These methodologies are described in the following sections.

2.2.6.1 Retrieval and ranking of textual data from the Internet

One of the uprising information needs in distribution systems analysis is knowledge about customers' willingness to participate in DSM actions, and conditions under which they would accept to change their daily habits in electricity usage. Next to the development of numerical data mining methods to retrieve this information (from surveys, for example), there is also an untapped potential among existing and publicly available textual data to find out these and similar pieces of information with the use of text mining methods. Although this data is usually given in tables or in the form of plain text in technical reports and research papers (offline), there is a possibility of using the Internet as a large database for extracting useful information.

There is a vast research carried out in different parts of the world about power networks and plenty of projects are focusing on the DSM perspective. Many distribution companies have selected pilot sites to carry out trials, aiming at investigating operation possibilities and consumers' opinion about DSM. Although these trials are not large-scale, the corresponding findings are capable of providing basic knowledge in all aspects of the DSM. Results of these projects are given in numerous reports and research papers, most of them being available on-line. This represents a valuable source of textual data accommodating different types of information about DSM.

In this methodology web crawling (traversing Internet to harvest documents of interest) and document ranking are combined to identify documents and sentences, using DSM as the topic of interest. The first step towards the discovery of useful information from textual data is to acquire adequate textual documents from Internet and then, in the following step, filter out less relevant parts of the documents to obtain meaningful information.

After generating a list of relevant search terms (key words) using either experts' knowledge or an automatic term recognition tool (e.g. FlexiTerm [140]) applied to a document highly relevant to the topic, the next step is to crawl webpages to obtain documents related to the desired topic, i.e., the list of keywords. Web crawl results with downloaded content of uniform resource locators (URLs) in the search engine (in this example, Google). As the main interest was on research papers and project reports, only web pages containing PDF and PowerPoint (PPT) documents were taken into account, since these on-line documents usually carry most information in the power networks area. Therefore, as the next stage, the resulting files were converted into text format as a necessary step before text mining.

As mentioned earlier, two parts of the methodology are explored: one to rank documents resulting from the web crawling process (using DSM as the topic of interest), and the other to extract important sentences, i.e., perform a short literature review (using power quality as the topic in this case).

Ranking of documents has an aim of finding documents that are highly related to a predefined topic in a semi-automatic way, with as little of user's interaction as possible. The generation of a ranking table of terms (keywords) is one of the most significant steps in this process, as the ranking of documents is based on the quantity of relevant terms they contain. There are three criteria combined to derive the final ranking of documents: the total number of occurrences of keywords (hits), weighting (importance) of the keywords (as shown in Table 2.2), and the percentage of text containing keywords in each document. Considering these criteria together, the final output is a ranked list of all documents according to their total score.

Table 2.2 Weightings of terms

Weighting	Relevance to the topic
5	Terms that are strongly related to the given topic
4	Terms that are not likely to be mentioned by other topics
3	Terms that could be mentioned by other topics
2	Generalised terms and concepts
1	Terms that are more likely to be mentioned by other topics

Sentence extraction from a series of documents aims to provide a summary on a given topic. A score is calculated for each sentence based on the number of keywords in the given sentence and the position of the sentence in the document. The first step towards the extraction of sentences is to present the frequency of terms (key words) in a form of a matrix called *Term Frequency – Inverse Document Frequency* (TFIDF) matrix, constructed as illustrated in equations (2.15-2.19) [141]. Each document d_i $(i = 1,..,m)$ in a set of m documents is presented as a vector of n terms:

$$d_i = (f_1, f_2, \dots, f_n) \tag{2.15}$$

where f_j $(j = 1,..,n)$ is the frequency of the term t_j in the document. Therefore, the term-frequency matrix for the set of m documents is given as a $m \times n$ matrix:

$$D = (d_1, d_2, \dots, d_m) \tag{2.16}$$

In order to eliminate the bias coming from long documents which as a consequence have higher frequency of the same term in respect to short documents, augmented normalised term frequency is used as follows [142]:

$$TF(t, d) = 0.5 + 0.5 * \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}} \tag{2.17}$$

where $f_{t,d}$ is the term frequency of term t in document d and $max\{f_{t',d} : t' \in d\}$ is the maximum frequency of any term in the document. Since the common words, such as "the", usually have high term frequency but do not carry any information, the inverse document frequency IDF for term t with respect to database D is defined to represent the number of documents within database D that contain t, as illustrated:

$$IDF(t, D) = ln \frac{m}{|\{d \in D : t \in d\}|} \tag{2.18}$$

where $|\{d \in D : t \in d\}|$ is introduced as *document frequency* (DF), representing the number of documents within the database that contain term t_j. Finally, $TFIDF$ matrix is obtained as follows:

$$TFIDF(t, d, D) = TF(t, d) * IDF(t, D) \tag{2.19}$$

The use of $TFIDF$ matrix will maximise the impact of a term when both conditions are met: high term frequency and low document frequency, i.e., small number of documents having high frequency of a specific term. Thus, due to the logarithmic characteristic of IDF function, the common terms tend to have extremely low IDF so that they are filtered out after the calculation of $TFIDF$ matrix.

In the following step, each sentence in the documents is given a score as follows [143]:

$$score = \frac{\sum_{t^i \epsilon s}^{m} TFIDF(t^i, d, D)}{\sum_{d^j}^{n} \sum_{t^i}^{M} TF(t^i, d) * TFIDF(t^i, d, D)} \tag{2.20}$$

where m is the number of terms in sentence s that is also included in $TFIDF$ matrix, M is the total number of terms (columns) in $TFIDF$ matrix and n is the number of documents in the database. In order to include the impact of heading (or the first sentence) as the indicator of the topic of a document, another score is added to the one defined in (2.21):

$$extra\ score = \frac{len(t)}{len(T)} * 0.1 \tag{2.22}$$

where $len(t)$ is the number of term/terms shown in both the target sentence s and heading/first sentence, and $len(T)$ is the number of terms shown in heading/first sentence and $TFIDF$ matrix.

The final step is to include the impact of sentence location, as adopted from [143]. The position P of a sentence is calculated as the line number of the sentence divided by the number of all lines in the text and presented in the range between 0 and 1. Based on Table 2.3 (adopted from [143]), another measure called distributed probability (DP) is obtained. The final score for each sentence is then calculated as follows:

$$final\ score = (score + extra\ score) * DP \tag{2.23}$$

Table 2.3 Distributed probability of important sentences (adopted from [143])

Position	$0 < P < 0.1$	$0.1 < P < 0.2$	$0.2 < P < 0.3$
Distributed probability (DP)	0.17	0.23	0.14
Position	$0.3 < P < 0.4$	$0.4 < P < 0.5$	$0.5 < P < 0.6$
Distributed probability (DP)	0.08	0.05	0.04
Position	$0.6 < P < 0.7$	$0.7 < P < 0.9$	$0.9 < P < 1.0$
Distributed probability (DP)	0.06	0.04	0.15

2.2.6.2 Case study 1: Document ranking

In order to discover customers’ opinion about DSM, a paper giving a literature review on this topic [144] was chosen to derive a list of keywords. After the automatic analysis of the paper using FlexiTerm, the terms and their corresponding term frequencies inside the given paper were ranked as presented in Table 2.4.

Finally, all documents from the database (PPT and PDF files) obtained through web crawling were ranked according to the level of importance to the defined topic. The titles of the first six papers and reports having the highest score are shown in Table 2.5. All highly ranked papers from the list showed to be highly related to the topic of both DSM and customers’ involvement in DR programs. They provide valuable information about users’ flexibility from both technical and economic aspects. Further reading of these documents would certainly give deeper insight into the subject.

Table 2.4 Key words for the given paper given by FlexiTerm

Rank	Term representative	Score	Frequency
1	literature review of major trials	152.97	142
2	domestic sector	111.36	163
3	critical peak	110.31	161
4	trial literature	97.98	145
5	literature review of major trials overview	97.53	72
6	literature review	97.34	144
7	peak demand	87.36	128
8	peak period	71.56	105
9	peak reductions	70.81	104
10	peak demand reductions	70.59	66

Table 2.5 Overall ranking of the documents

Ranking of document	Paper title	Number of pages
1	"Demand side response in the domestic sector – a literature review of major trials"	156
2	"Residential Demand Response for outage management and as an alternative to network reinforcement"	64
3	"Smart Tariffs and Household Demand Response for Great Britain"	93
4	"Assessment of Demand Response and Advanced Metering"	92
5	"Developing the smarter grid: The role of domestic and small and medium enterprise customers"	50
6	"The role of demand response in electric power market design"	57

2.2.6.3 Case study 2: Sentence extraction

The second case study represents sentence extraction methodology applied to the topic of 'power quality'. The database in this case consists of 783 documents obtained via web crawl. With the methodology for allocating score to each sentence, a text file containing 30 sentences with higher scores within the database is generated. For illustration purposes, there are four sentences with the highest scores presented in Table 2.6 to show the quality of extracted sentences.

In these four sentences, it is stated that power electric utilities are trying their best to maintain high standard of power quality, especially in open market, and thus they have to take into consideration the impact of harmonics and conductive disturbances in voltage supply. At the moment, the information provided by these highly ranked sentences is not logically and semantically arranged, since it is still at the stage of sentence extraction. However, this tool is capable of providing an overall idea of the topic after reading all extracted sentences. Another option for the user in this case is to track back from the extracted informative sentence to its original source (paper) and obtain more detailed information without having to read other documents in the database.

Table 2.6 Extracted sentences

paper 322	Conductive disturbances in the supplying voltage may also affect the root mean square voltage and shape of the voltage curve on the mains which reflect in degradation of the power quality.
paper 342	Introduction: The power quality problem is now of a great concern to electric utilities of power industry and they are trying hard to supply their customers with a good quality of power especially in the open market.
paper 71	Conclusion: the simulation approach provides the researcher the flexibility to create power system models to simulate power quality disturbance by connecting various functional building blocks in the simulation environment.
paper 342	Today, a new factor, harmonics, has been added to the power quality scenario because utility customers, including residential ones, are using electronic devices that require non-sinusoidal currents, currents rich in harmonics.

2.2.6.4 Summary on text mining methods

The previous subsections have shown that, although still underutilised, text mining methods have a great potential for applications in power system studies. Ranking of documents, which may be customer surveys or technical reports, could facilitate overview of world-wide practices in an area. This can further pave the way to changes in regulatory and market frameworks, for example. Sentence extraction, resulting in a summary on a certain topic, could bring similar benefits by extracting key information from documents or from textual data coming from social media. Potential for combining textual data from different sources, such as reports, Internet and social media, is yet to be discovered, and this can be achieved in future by collaboration between experts in power systems and text data analysts.

2.3 Data needs in future distribution networks

As mentioned earlier, one of the most important distinguishing features of future power grid operation will be the increased use of ICT in the generation, transmission and distribution of electrical energy [145]. With the complexity of modern distribution power systems grows the size of their monitoring systems and databases containing variety of data coming from numerous monitors and sensors. Databases are increasing in two dimensions: in the number of objects (instances) and in the number of fields for attributes describing those objects [114].

Distribution system utilities' servers are constantly receiving and storing large amount of real-time data keeping enormous memory busy with numbers and text. It has not been analysed yet how thoroughly these dynamic data are being processed, i.e., how much knowledge has, and can be retrieved from the existing collection of data. Similarly, a significant amount of static data is contained in tables and reports, given in numerical and textual form. As it is stated in [146], "a very large amount of data is being collected whose potential has been untapped".

Dynamic data sets in distribution network most commonly involve measurements of real and reactive power, voltage, as well as power quality (PQ) measurements (an example of monitored parameters is shown in Table 2.7 [147]).

Table 2.7 Power quality parameters and corresponding time scales (adopted from [147])

Parameter	Timescale
Voltage transient	<20 ms
Voltage dip	10 ms to 2 s
Frequency excursion	Possibly one minute
Phase unbalance	Possibly one day
Harmonics (percentile)	1 week
Flicker (percentile)	1 week
Dip/swell statistics	3 years

Electricity consumption data is collected from conventional meters or SMs at end-users' premises, and balancing meters at distribution substations [148]. Depending on the source, data is given in various data ranges, i.e., in milliseconds (from phasor measurement units - PMUs), seconds (from SCADA systems) or minutes (from SMs) [132].

Power distribution utilities' databases are, as many other types of databases, characterised by several common features [149]:

- Large size;
- Noisiness;
- Incompleteness or absence of records;
- Semi-random survey design (redundancy of records of one variety but a lack of records of another);

- Enormous heterogeneity of response variables and large number of predicting variables.

Much of the collected data from the distribution grid has not been used at all, i.e., it has not been transformed into knowledge. The main problem coming from the installation of new monitoring systems is the availability of memory space in database servers and handling already existing and newly coming data. One of the solutions to these issues would be finding hierarchy in the importance of data, i.e., finding redundant data types and excluding them from monitoring systems. In addition, granularity of data, i.e., the sampling step of measurements, also affects the required memory size, which is why there should exist a trade-off between the granularity and the usefulness (application) of the data being collected. Another challenge is prediction of the types of information future (smart) distribution grid operators will need in a time frame of 10-20 years. Therefore, it is very important to perceive future electricity market actors, their functionalities and assets that will be used. This may disclose possible additional types of data that will have to be collected. In that respect, an overview is made of the types of data that distribution utilities are already collecting.

The existing substation monitoring devices are most commonly collecting the following data:

- Basic measurements: voltage magnitude, real power, reactive power, power factor, apparent power, phase sequence, voltage phase angle, current phase angle, neutral current;
- Power quality data: voltage sags and swells, harmonic distortion per voltage and per current phase, total harmonic distortion;
- Asset status: power transformer tap position, breaker status [150];
- Faults data: number of outages, date and time of the last outage, cumulative power outage time.

One of the main features of future distribution grid will be the extensive collection of electricity consumption data from the end-users via SMs. The aim of the smart metering system is to make power metering two-way, so that electricity suppliers could pass on the dynamic (daily and seasonal) change of electricity price to customers, which would incentivise them to save energy and make savings to both themselves and suppliers [75].

In addition to SM data, there are some complementary types of information deemed as necessary for operating the distribution network in smart grid environment, such as:

- Home occupancy, i.e., how many household occupants there are, and how much time they spend at home, consuming electricity [151]. Customers' needs, daily routines and lifestyles are very valuable pieces of information as they can be correlated to the electricity consumption. These types of information have been collected through customer questionnaires;
- Operational characteristics of devices under different environmental conditions – this is useful for the investigation of potential types of appliances that might be controllable now or in the future as a part of DR programs;
- The amount of controllable loads in a particular area - automatic control of the load is still an aim for the future, although some utilities already operate direct load control programs, for example by sending a radio signal to water heaters to automatically turn them on and off [152];
- Net consumption, i.e., net metering for consumers who own renewable generation facilities – this is either in use already (as in the Netherlands, Sweden, Italy, Hungary, UK, Finland and Denmark), or is planned to be introduced (in Croatia, Cyprus, Estonia, Greece and Romania) [11];
- The average installed power per household in an area – this is very important with respect to implementation of modern appliances, such as electric vehicles or heat pumps. If the existing electrical installation is not designed for the increase of load, the size of the additional investment plays a big role in decision making. The UK government estimates that the country could have 20 million new heat pumps installed by 2050. Similar scenario could happen in the rest of Europe (except in its southern part), where space and water heating account for a big share (usually more than 75%) of home energy consumption [153];
- Weather conditions – except for their influence on the daily load profile of the consumers, this information could also enable DR programs which could shift demand to windy or sunny times of the day, when the renewable energy sources are generating the most [154];

- Statutory voltage limits for the supply of power to customers [154]. In the UK, these limits are between -6% and 10% for 230 V, while phase imbalance should not exceed 1.3% [155];
- Capacity of the circuit to carry power [154] - several variables need to be considered, including the thermal capacity of network elements (at both transmission and distribution level), the need for operating reserve for stability reasons (especially at transmission level) and permissible voltage variations (especially at distribution level) [11];
- Possible network topology changes and longer-term change in demand [154];
- Number of interruptions in an area [154].

The state-of-the-art information needs in distribution network are summarised in Table 2.8. Six areas of interest (knowledge) are presented, together with the appropriate groups of data that are already used or might be used additionally to obtain valuable information.

Data types that have to be forecast are weather, demand, available renewable energy, state of charge of storage units and price. In addition to the types of data stated in group 1, it would be highly useful to obtain data about the percentage of customers capable of participating in direct load control. Among the controllable loads, of special interest are electric vehicles and thermostatically controlled loads (HVAC, heat pumps, water heaters, freezers and refrigerators). Also, location of the DR resources is important for the power flows and potential “matching” with the available distributed generation. In group 2, information that would be of interest is the percentage of distributed generation and storage devices that can be controlled as dispatchable resources by the DSO [145]. Smart meter measurements in group 3 are of particular importance, as to this point distribution utilities do not have any knowledge about outages unless they receive calls from customers. The percentage of grid assets that are monitored, controlled, or automated could be added to data in group 4.

As an example of the types of data requirements that can presently be met solely from surveys, but that could in the future be provided by monitoring systems, Figure 2.6 illustrates the DSM potential (flexibility) in transmission networks in South-East Europe [156]. The potential is given as capacity in MW, and in the number of potential end-users (large industrial users connected directly to the transmission network) that could provide flexibility.

Table 2.8 State-of-the-art information and data needs in distribution network analysis

Required information (knowledge)	Data collected	Sampling step (commonly used)	Data group
Short-term demand forecasting and flexibility assessment (including demand decomposition) at primary substation level/aggregator level to facilitate the process of network balancing and demand response	Voltage, real and reactive power	1—30 min	1
	Weather data	30 min—1 h	
	Pricing data	15 min—1 h	
	Customer's affinity to different types of incentives (monetary or not)	/	
	Percentage of customers with smart metering (including sub-metering) – observability level	/	
	Real and reactive power of controllable loads	1—30 min	
	Customer's willingness to participate in demand response	30 min—1 h	
Short-term distributed energy resources forecast at primary substation level/aggregator level to facilitate the process of network balancing and demand response	Voltage, real and reactive power	1—30 min	2
	Weather data	30 min—1 h	
	Amount of distributed energy resources (PVs, small wind turbines, storage)	/	
	State of charge of storage units/electric vehicles	1—30 min	
Fault location identification (fault detection)	Current and voltage waveforms	~ms/s/min	3
	Smart meter measurements	1—30 min	
	Network topology	/	
	Historical faults data	/	
	Relays and breakers states	~ms	
	Power outage time	/	
Condition assessment / Asset management	Age of assets Monitoring data acquired during the operation Number of faults Network topology Line capacity	/	4
Electricity price construction and price spikes forecast	Electricity market data	15 min—1 h	5
	Historical data on consumption	30 min—1 h	
	Weather data	30 min—1 h	
Power quality detection for power system disturbance	Voltage and current waveforms	~ms	6

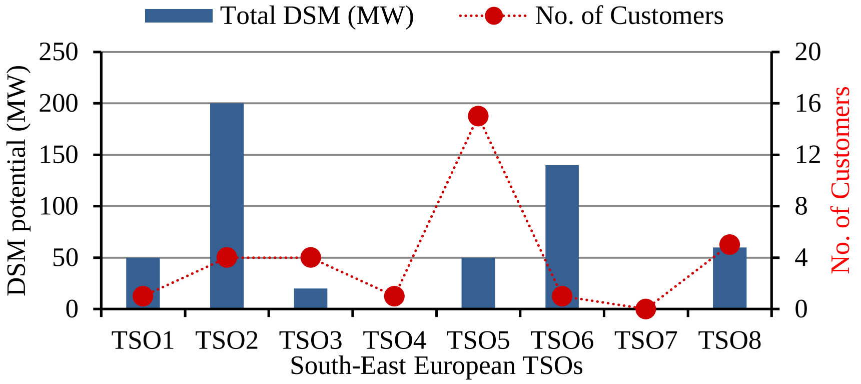


Figure 2.6 DSM potential in eight TSOs in South-East Europe (adapted from [156])

Following this, the survey identified types of industrial users with the highest DSM potential in the observed region, as shown in Figure 2.7. Availability of this type of information facilitates more accurate load modelling of the DSM providers, and enables more accurate steady-state and dynamic analysis of the power network and the effects DSM could introduce in the network. With appropriate smart meters or similar sub-metering technologies installed at the premises of these end-uses, one could observe and follow the daily changes in demand and its composition, i.e., demand flexibility.

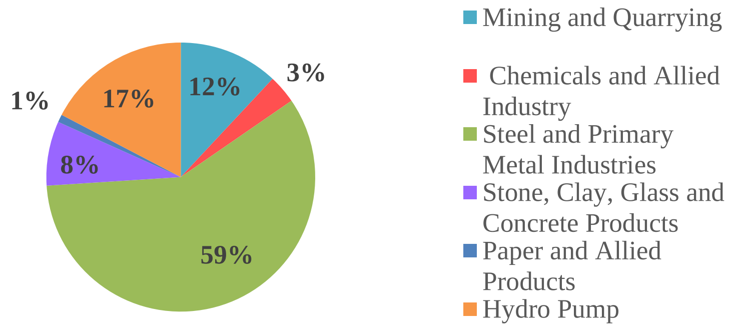


Figure 2.7 Largest DSM providers in South-East Europe (adapted from [156])

2.4 Smart metering versus future distribution network data requirements

There are numerous sources giving information about the benefits SMs will bring to the end-users. On the other hand, more insight is needed into the benefits provided to the network operator, at distribution and transmission level. Following a wide scale deployment of SMs, customers will benefit from the following advantages [157, 158]:

- Near real-time information and updates on energy use, the cost and carbon-dioxide (CO_2) emission;

- Better management of energy usage, saving money and reduction of CO_2 emission;
- Billing based on the actual consumption, not estimation;
- Easier switching to other suppliers with different tariffs;
- Access to historical consumption data.

The expected benefits for the network operator are the following:

- Monitoring of low-level consumption, which facilitates more accurate load forecasting at the distribution level;
- Faster identification of faults and users causing non-technical losses (fraud);
- More accurate consumer profiling for tariffing purposes and DSM;
- Facilitating load reduction and load shifting [159];
- Financial savings (it was reported in [160] that utility companies are expected to save $157 billion by 2035 by using SMs).

2.4.1 Smart meter specifications

In the UK, customers have the freedom to choose whether they want the measurement data to be sent monthly, daily or every 30 minutes [158]. This, however, will cause difficulties to applications requiring the same sampling step of the incoming data streams which are to be aggregated and further processed. Another issue may be the lack of synchronism between distributed data streams coming to the same concentrator (aggregation) point.

The meter should be recording active and reactive energy import and export and keep all the information in its own data store [161]. Based on specifications given in [162-164], SM accuracy of measurements complies with class 1 (error limits ±1.5%) or class 2 (error limits ±2.5%) for active power/energy and class 2 for reactive power/energy. There should be channels for load profile recording, as well as for appliance profile recording. Profiling period can be between 1 and 60 minutes, or one day.

According to technical specifications required by the UK Government's Department of Energy and Climate Change (DECC) [161], SMs should also measure average root mean square (rms) voltage and record cases of over or under voltage (i.e., when the value is over the 'average rms over-voltage threshold' or under the 'average rms under-voltage threshold', respectively). It should also detect voltage sags and swells. There should be a load switch for enabling/disabling supply, and limiting power consumption. Data storage is required to keep minimum of 13 months of active energy imported and 3 months of active energy exported and reactive energy imported and exported. The technical requirements, however, do not mention how the SMs will be monitoring smart home appliances in the future, or how the direct control of appliances will be actualised.

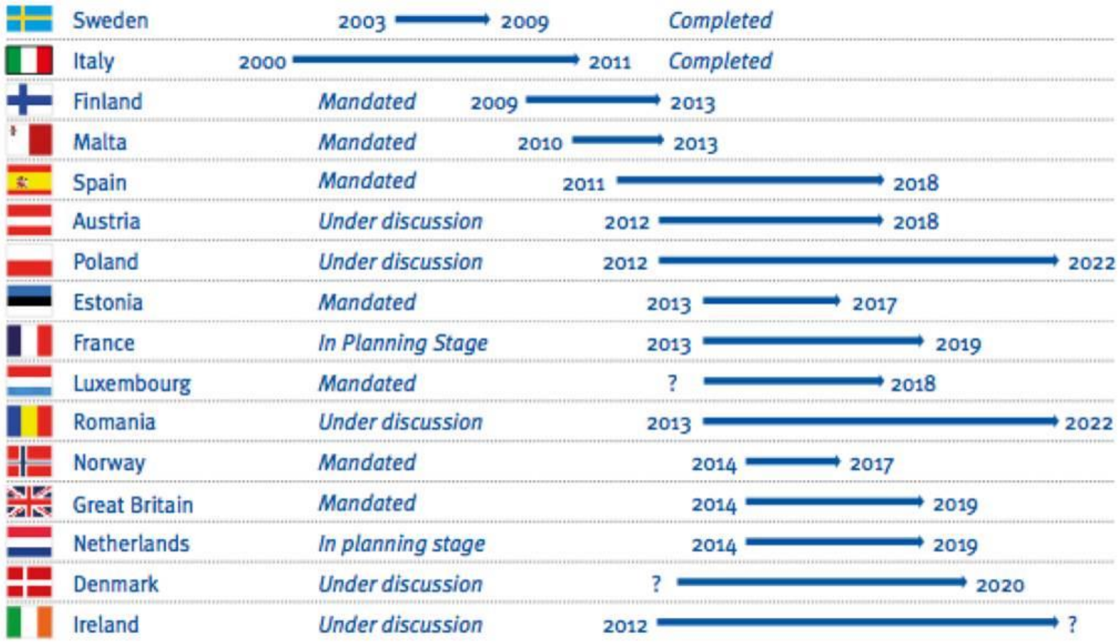
The commonly used communication systems for smart metering are Radio Frequency (RF) technology and General Packet Radio Services (GPRS) systems. The best known RF architecture is RF mesh, where SMs form a Local Access Network (LAN) which, in cases when a node fails (drops out of the network), enables the signals to find another route via the active nodes [165].

The European Union (EU) specifies the following requirements for different aspects of the smart metering system [166]:

- The consumer: readings to the consumer and/or a 3rd party should be enabled and updated frequently enough to allow energy savings. The recommended update rate (sampling step) is 15 minutes;
- The metering operator: remote and frequent reading should be enabled, as well as a two-way communication for maintenance and control;
- Commercial aspects of supply: smart metering system should support advanced tariff system and allow for remote on/off control of supply and/or flow or power limitation;
- Security and data protection: the system should provide secure data exchange and enable fraud prevention and detection;
- Distributed generation: the system should enable real power net metering and reactive power metering.

2.4.2 Smart meter rollout in Europe

It is foreseen that around 50 million smart electricity and gas meters will be installed in the UK by 2020 [167], out of which 27 million smart electricity meters in the domestic sector [4]. The rollout started in 2015 costing an estimated £11 billion, and is expected to deliver a net benefit of £6.7 billion [168]. The six main energy suppliers in the UK are leading the rollout, coordinated by government with industry support, while the SM data is the responsibility of a recently founded Data and Communications Company (DCC). The roll-out is being coordinated by the Department of Energy and Climate Change (DECC), and, once the meters are in place, the program will be governed by the national regulatory authority – the Office of Gas and Electricity Markets (Ofgem). In most European countries the target is to have at least 80% of the end-users with SMs [166], as illustrated in Figure 2.8. That represents an estimated number of about 195 million SMs and a total investment of about €35 billion.



Sweden		2003	2009	*Completed*
Italy		2000	2011	*Completed*
Finland	*Mandated*	2009	2013	
Malta	*Mandated*	2010	2013	
Spain	*Mandated*	2011	2018	
Austria	*Under discussion*	2012	2018	
Poland	*Under discussion*	2012	2022	
Estonia	*Mandated*	2013	2017	
France	*In Planning Stage*	2013	2019	
Luxembourg	*Mandated*	?	2018	
Romania	*Under discussion*	2013	2022	
Norway	*Mandated*	2014	2017	
Great Britain	*Mandated*	2014	2019	
Netherlands	*In planning stage*	2014	2019	
Denmark	*Under discussion*	?	2020	
Ireland	*Under discussion*	2012	?	

Figure 2.8 Smart meter rollout in Europe (adopted from [165])

The largest expected number of installations is in Italy (36.7 million), France (35 million) and UK (32 million). Due to the negative outcome of the cost-benefit analysis (CBA), some countries opted out from wide-scale rollout (e.g., Belgium, Lithuania and the Czech Republic). In countries such as Germany, Latvia and Slovakia, the CBA outcome was reported negative for a large scale rollout, but economically justified for a specific group of customers. Even though there are already many SMs installed in residential properties throughout Europe, for the time being there is no data available from them, except for a small number of trial installations in pilot sites [154].

2.4.3 Benefits of smart meter data

Compliance between the present SMs' features and data they provide on one side, and information needs for operation of the future distribution network on the other side are presented in Table 2.9. As illustrated in the table, most of the present or foreseen data needs can be, to smaller or larger extent, met by SM data. The data requirements that are still not covered by commercial SMs are indicated by shaded cells in the table.

Considering the potential use of data coming from SMs, the following system/operator functionalities/actions would be greatly enhanced by the use of SM measurements:

- Load forecasting could be applied to lower levels of aggregation, because it would follow daily pattern of customers in a specific distribution network area. This could further facilitate local DSM programs.

Table 2.9 Smart meters features and their compliance with future DSO's needs

Smart Meter Features		Information and Data Requirements					
		Load forecast	Fault detection	Distribution network state estimation	Amount of controllable loads	Customers' willingness to participate in DSM	Amount of DER
Measurements	imported active power/energy	X		X	▨	▨	
	exported active power/energy				▨	▨	X
	imported reactive power/energy	X		X	▨	▨	
	exported reactive power/energy				▨	▨	X
	rms voltage	X	X	X	▨	▨	
Detection	under voltage		X		▨	▨	
	over voltage		X		▨	▨	
	voltage sags		X		▨	▨	
	voltage swells		X		▨	▨	
Sampling step	15 min (Italy)	X	X	X	▨	▨	X
	30 min (UK)	X	X		▨	▨	X
	60 min (Sweden)	X	X		▨	▨	X
Additional features	load switch		X		▨	▨	

- The accuracy of state estimation could be largely enhanced with the reliance on low-level consumption (real and reactive power) data, which are presently hardly accessible and hence replaced with pseudo-measurements. State estimation algorithms will require near real-time (within several minutes) voltage measurements to allow active voltage control, especially in cases of unpredictable load profiles in presence of DR, electric vehicles (EV) and heat pumps (HP). If no real-time data are available, distribution network voltages can be reliably estimated using SM data from the previous-day [167].

- Fault detection could be highly improved with smart metering of voltage, enabling the network operator to receive notifications about interruptions in supply much faster than usual (distribution network operator commonly receives fault notifications from customers service). This would drastically reduce the supply restoration times.

- Since the majority of presently installed SMs do not have the functionality of monitoring individual appliances, estimation of the amount of controllable loads can at this point be made only by using some statistical data and probabilistic approach. On the other hand, there are a number of pilot sites with SMs measuring consumption of individual appliances. Therefore, it is not far from reality to assume that this option might become common in the near future, especially with the increasing number of new smart home appliances in the market. Per-appliance monitoring could enable observability of the end-users' flexibility (in real-time or in the future), as well as real-time monitoring of the obtained DR.

- The information about customers' willingness to participate in DSM actions would reduce uncertainties about the actually available load flexibility in cases of voluntary DR contracts between customers and the network operator, where customers react to DR (i.e., incentives to shift or reduce their consumption). Although SMs do not have this feature, signals about confirmation of participation in DSM could also be sent through other devices, e.g., smart phones or personal computers.

- The amount of flexibility offered by distributed energy resources (DER) could also be "tracked" thanks to SMs measuring energy flow in both directions, i.e., both import and export.

Some authors [30] stipulate that SMs also need to embed some level of local data analysis capabilities, since sending huge datasets (for millions of customers) along the communication channels exceeds the current capacity of these channels. In addition to this, SMs are still not considered to be suitable for on-line monitoring of networks for control purposes because of the high data latency (time needed to collect and store data) of the national communications system which is often chosen to collect SM data [154].

Another type of information that might potentially be provided via SMs is the total capacity of the supplying cable in households, which would be necessary for planning deployment of any new technology such as HP or EV.

2.4.4 Challenges arising with smart meter technology

Even though the deployment of SM technology offers significant advantages in terms of improved accuracy of various functionalities needed for DSM, there are still some issues that need to be resolved. Since the customers will decide on the frequency of sending their load profiles to the data concentrator, data streams will not be coming with the same time steps. This will result in additional missing data, considering that some data might be missing anyway due to malfunction of devices or communication failures. Therefore, there will be a need for development of off-line or on-line data restoration methods, depending on the application. Influence of this data pre-processing on the accuracy of aggregated data is yet to be investigated. Furthermore, considering that all measurements contain a certain amount of noise, the appropriate filtering of measured signals is required. The level of filtering and accuracy of filtered data will depend on the type of application that the data will be used for, so adaptive filtering techniques need to be developed.

Operational challenges such as software and hardware faults, and malfunction of SMs present a realistic impediment to successful and timely data aggregation [169]. Possible technical issues include intermittent communication networks, insufficient signal strength and inability to detect a communication network failure. These may also lead to missing data and network latency, and hence aggravate the problem further.

For on-line applications, such as state estimation and DR, higher granularity of data is needed (e.g., minute based), so 30 or 60-minute based sampling steps that are presently most widely used, may not be appropriate. Also, in order to reduce the computation time, it is preferable to receive power data from SMs (averaged over small time steps), instead of energy data over a time frame. Finally, for any type of transient stability analysis that will include studies of dynamic response of demand [16], granularity of power samples should be even higher, in the range of seconds or even milliseconds. In this case signal latency and synchronised sampling become an important issue, in addition to significant increase in amount of data that would need to be processed.

As reported in [167] there are still unresolved questions regarding the rollout and future use of SMs, among others:

- How can SMs be used to report and verify DR effect?
- Can SMs be used to precisely locate failures at individual consumers' premises?
- What techniques can be used to analyse SM data?
- How to balance between consumers' privacy and use of data by third parties?

Furthermore, there are also safety risks coming with application of SMs, in particular cyber-attacks, which may result in remote disconnection of a large number of customers and changing the loading of the power network, which would affect the reliability and security of the system as a whole.

2.5 Data analytics methods in distribution network analysis

Distribution system operator (DSO) controls a much larger number of power lines and substations than a transmission system operator (TSO), which makes distribution network less observable. UK LV network, for example, involves 230,000 HV/LV substations, including 580,000 transformers and 376,000 km of overhead lines and underground cables [170]. A lot of effort has been put to ensure the optimal control and operation of the distribution system despite the reduced observability. Hence, the use of data mining methods is taking the lead as a cost-effective means of gaining additional and useful information from raw data arriving from a limited number of monitoring devices.

To ensure the effective use of the acquired data, a system for collection and processing of the data coming from power system monitoring devices and databases should have the following characteristics [171]:

- Flexibility for storing data coming in different forms from various sources, such as Distribution Management System (DMS), Automatic Meter Reading (AMR), Excel files, text files, etc.;
- Flexibility in dealing with both static (network topology, technical characteristics of assets) and dynamic data (on-line and off-line measurements);
- Automatic calculation of performance indicators;
- User-friendly interface for reading, editing or managing data;
- Extensibility for future needs and other forms of incoming data [119].

Classification and clustering methods have already been widely used in distribution system analysis for grouping individual customers with similar electrical behaviour (load pattern) [29] or for grouping feeders with similar features in the network [31]. The former can be applied for “tailoring” the tariffs and DSM programs for different classes of end-users or for bad data identification, more accurate demand forecasting and network planning purposes. Another application is detection of the penetration level of low carbon technologies (LCT) at the demand side (EVs, heat pumps, storage, renewable generation) by detecting changes in the baseload of the end-users [172]. The latter can be used for facilitating network maintenance or for assessing the hosting capacity of the numerous feeders in the network. Furthermore, different load buses in the distribution network may have similar composition of load types, similar daily or seasonal load patterns, similar load location, and so on. Classification or clustering techniques bring the possibility of clustering buses into groups based on certain features and monitoring only one representative substation (bus) in each group. In this case, the load model generated based on load monitoring at the representative bus can be applied to all buses assigned to the same group. This drastically reduces the cost of implementation of monitoring systems [173]. In [170], LV substations were classified according to their location and customer dominant type information, including population, consumption and economic situation.

Load pattern classification is commonly done using k-means clustering or its variations (e.g., fuzzy k-means, where samples may belong to multiple clusters, but with a different degree of membership), hierarchical clustering or self-organising maps (a type of unsupervised neural networks) [135]. When clustering daily load patterns, the main steps of the process are as follows:

1) Removal of bad data or noise;
2) Categorisation of measurements based on the type of the day (season, working/non-working day);
3) Normalisation of the data for comparability;
4) Clustering;
5) Extraction of the representative pattern for each cluster (these are usually the centroids obtained at the end of the clustering process).

k-means clustering has also been used for missing data restoration in load profiles measured by SMs [169]. Segments of load profiles were clustered based on their similarity, and the most similar cluster centres were used to restore the missing samples at the corresponding time steps.

Correlation analysis is highly useful for identifying attributes of data which can be used as predictors of the unseen data (whether it is missing data or data that has to be forecast). For example, due to the high correlation between weather (most commonly the outside temperature, humidity and wind speed) and electric load, weather forecast is used to predict demand based on a model trained with historical data measurements of demand and weather. As another example, correlation analysis of voltage profiles at network buses was used in [174] to identify those buses that are most sensitive to disturbances. Authors in [175] allocated daily load curves of the predefined (sample) end-users to the end-users without a SM based on the correlation between the variation of their monthly energy consumption and the variation of the monthly energy consumption of the sample end-users. Correlation could also be used to analyse connection between SM data and other types of data coming from external databases, such as weather data, traffic data, social events (such as concerts or large sport events), etc., to enable more accurate demand forecasting or facilitate more effective DSM programs. In this case dynamic data processing or event processing [176] can be applied, processing events from different sources (meters, sensors, Internet, etc.). The aim of this processing is to detect event patterns and give detection or prediction of complex events, that otherwise would not be predictable.

Decision trees have been applied in power system stability studies using data from phasor measurement units (PMUs) [177]. They have also been used in fault detection, customer classification and estimation of energy usage [178].

ANNs are mostly used for load forecasting, stability and security analysis, power system control, fault diagnosis, reactive power planning and control and for state estimation [126, 179, 180]. As already mentioned, SOM, as a type of ANN for pattern recognition, are used in load classification [181, 182]. ANNs are also applied in non-intrusive load monitoring, for classification of appliances based on their current harmonics [36]. The deep learning method, as an upgrade of ANN which has been successfully applied in computer vision for object recognition, has recently been used for load forecasting of individual end-users [183] and for power quality studies [184].

2.6 Case study: data analytics methods applied to a distribution utility database

As an illustration of information retrieval from databases, statistical analysis and some of the main data mining methods are applied to a real distribution utility's SQL database with static data about faults on feeders in HV (6.6 kV and 11 kV) and LV network. The database consists of numerous tables showing feeder characteristics, i.e., feeder type, district, exact location of primary substations (33 kV/11 kV and 33 kV/6.6 kV), number of connected customers, etc. Also, given are exact dates and times of faults followed by the number of interruptions and cumulative duration of customer interruptions per fault. Number of customer interruptions (CI) and customer minutes lost per customer (CML), as the key indicators of quality of service (QoS) of the distribution network, are calculated as follows [185]:

$$CI = \frac{number\ of\ interruptions}{total\ number\ of\ supplied\ customers} \cdot 100\ \% \tag{2.24}$$

$$CML = \frac{cumulative\ interruption\ duration}{total\ number\ of\ supplied\ customers} \tag{2.25}$$

The data are aggregated to a five-year period, i.e., all the records show QoS performance from 2007 to 2011. Aggregation and connection of data was performed using SQL database queries. Datasets were further analysed in Matlab and Weka [186], which is an auxiliary and convenient tool for data mining and presentation. Weka was chosen for its simplicity of application – after the input data is uploaded in the form of a table (.csv file), the tool instantly gives graphical view of the statistical analysis. It also gives access to SQL databases [114].

The analysis is divided into two parts – HV analysis (6.6 kV and 11 kV feeders) and LV analysis (feeders up to 1 kV). Further aggregation of data was done at the voltage level (dividing HV network data into 6.6 kV and 11 kV data), feeder class (according to Ofgem classification of HV feeder types [187]) and district level (in order to compare key indicators performance among geographical districts). As the considered database is not big enough to justify the use of ANN or k-means clustering, linear regression and decision trees are applied in this case study.

2.6.1 HV analysis

An overview of QoS indicators performance at HV level is done following two approaches: by feeder class and by geographical district. The data were pre-processed, i.e., cleaned by removing instances with misleading values. For example, an instance could refer to a fault happening on a feeder not supplying any customers, but having a number of customers affected by the interruption. Even though it is possible to have a feeder whose interruption would consequently influence customers that are not directly connected to the feeder, these examples would lead to overestimated QoS indicators, such as customer interruption duration, which is why they were excluded from the analysis.

All HV feeders were classified into 11 classes based on the percentage of the overhead line (OHL) part with respect to the total length (in km) of a feeder, as well as the number of customers supplied by the feeder (Table 2.10). It should be noted that the word “All” in the table refers to all feeders, regardless of the number of customers connected to it.

Table 2.10 Characterisation of feeder classes

	UG1A	UG1B	UG2A	UG2B	MA1	MA2	MB1	MB2	MC1	MC2	OH
% OHL	0				<20		20-50		50-80		>80
Length (km)	<4		>4		<8	>8	<11	>11	<19	>19	All
Number of customers	<1000	>1000	<2000	>2000	All	All	All	All	All	All	All

2.6.1.1 Quality of service analysis based on the class of the feeder

The average cumulative interruption duration is shown in Figure 2.9, as well as the average number of interruptions per feeder and average number of faults per feeder, based on feeder class. Following the horizontal axis in the figure, the OHL length increases. Letters A and B in the UG class names refer to smaller and bigger number of customers connected to the feeder, respectively. It can be observed that the number of faults per feeder increases with the length of the OHL part of the feeder, as well as the total length of a feeder, while the average duration of interruptions and number of interruptions per feeder grow with the number of connected customers.

2.6.1.2 Cost of compensation for the energy not supplied

To make an estimation of the cost of compensation for the energy not supplied (ENS) to customers during interruptions, the value of lost load (VoLL) is adopted from [188]. With the assumption that the ratio of domestic customers and small and medium sized businesses (SME) is 74:26 ([188]), and that on a peak winter workday VoLL is around 10,000 £/MWh for domestic and 35,000 £/MWh for SME users, a load share-weighted average VoLL is taken to be around 17,000 £/MWh, i.e., 17 £/kWh. The VoLL for industry and commercial (I&C) customers is much lower than for SME (around 1,400 £/MWh), since these customers use more energy and are more likely to have self-supply [188].

Due to the complexity of assessing typical industrial consumption, only cost of compensation for the domestic and SME users is estimated in this study. In other words, only 11 kV feeders' data were taken into account for the cost analysis. QoS parameters were estimated according to both 6.6 kV and 11 kV feeders. The performance of QoS indicators over the 5 year period is illustrated in Figure 2.10 based on the feeder class, together with the cost of compensation for ENS. The values are normalised for comparison using base values defined in Table 2.11.

The calculation of the cost of compensation was done as follows:

1) The number of CML was calculated for every fault and multiplied by the number of interrupted (affected) customers;
2) The value calculated in 1) was then multiplied by the average domestic consumption, standing for 1.1 kW [189], giving the ENS per fault. This

value was summarised for all the faults happening on each of the feeder classes;

3) The cumulative ENS was multiplied by the value of lost load (VoLL) for domestic sector (16.94 £/kWh [188]), giving the compensation cost for each feeder class during the given period.

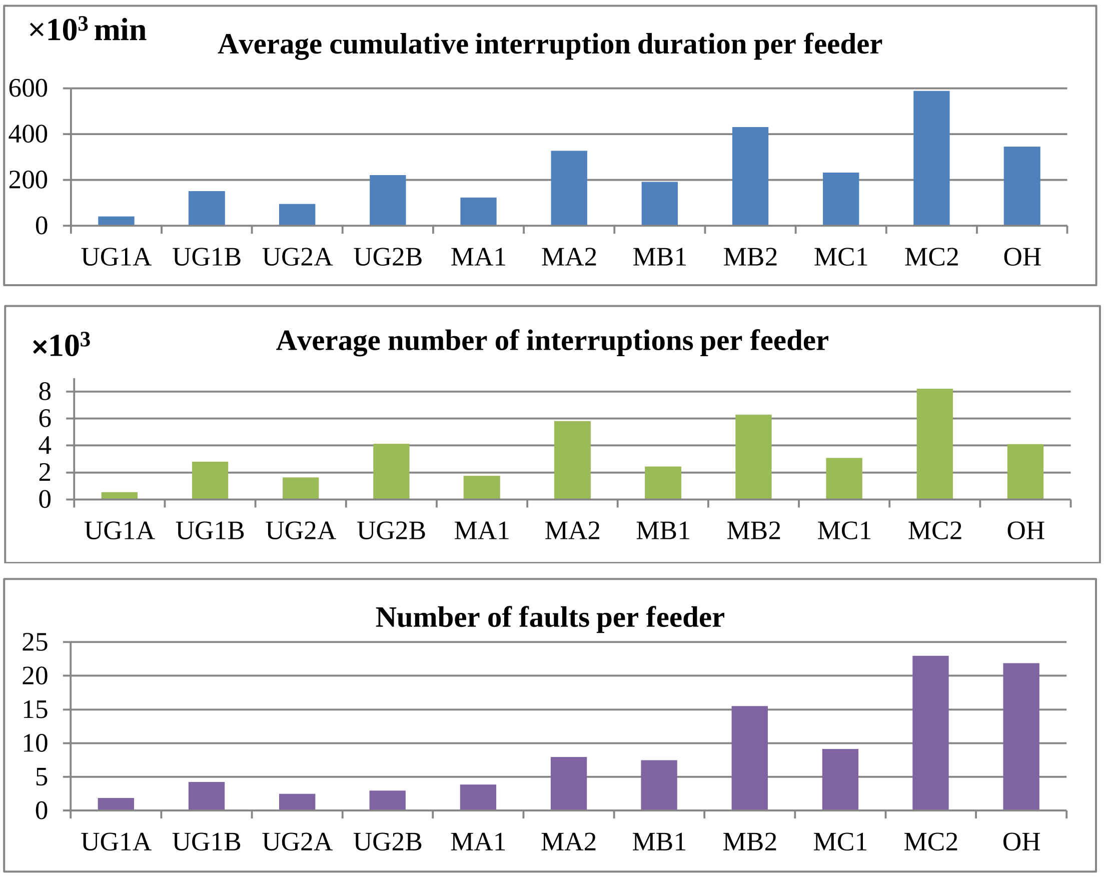


Figure 2.9 QoS indicators according to feeder class

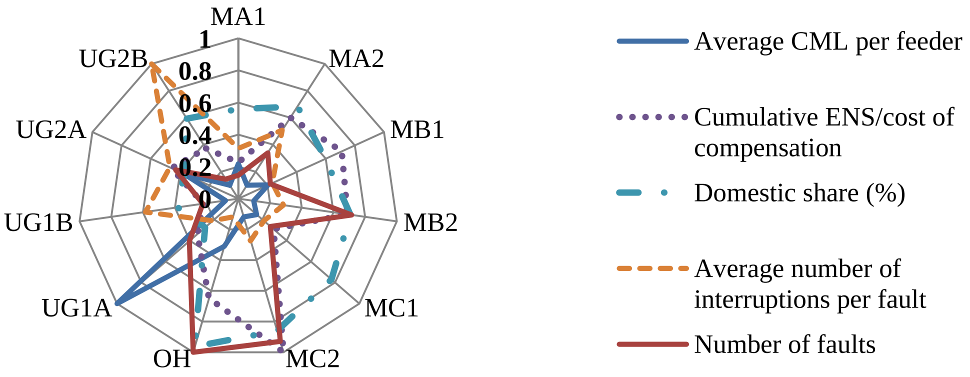


Figure 2.10 QoS indicators performance with normalized values per feeder class

Table 2.11 Base values for normalization in Figure 2.10

Measures	Base values
Average CML per feeder	18,860 min
Cumulative ENS/Cost of compensation	720 MWh / £12.2 million
Domestic share	100 %
Average number of interruptions per fault	1,445
Number of faults per feeder class	2297

As it can be seen from the radar diagram, the highest cumulative amount of ENS was calculated across HV feeders of class MC2 – feeders with a high share of domestic users and also a high share of OHL part. Following this, MC2 showed the highest compensation cost rate for domestic users during the period (around 1.8% of the five-year profit of the DSO managing the network). Feeder class UG1A (underground cable) showed the highest rate in average number of CML per feeder, probably due to the reduced accessibility for fault removal. This is also the most common type of feeder used in the observed HV distribution network, with a contribution of around 27%. Feeder class UG2B showed the highest number of interruptions per fault, which is justified by the fact that this feeder class supplies more than 2,000 customers on average.

The range of possible cost of compensation for ENS for domestic users, depending on the time of the outage, and the use of WTA (willingness to accept payment if an outage occurs) model are adopted from [188] and given in Table 2.12. Similarly, the range of VoLL for SME users is given in Table 2.13.

Table 2.12 Range of VoLL for domestic users [188]

	Other seasons				Winter			
	Off-peak		Peak		Off-peak		Peak	
	Weekend	Weekday	Weekday	Weekend	Weekend	Weekday	Weekday	Weekend
WTA (£/MWh)	9,550	6,957	9,257	11,145	10,982	9,100	10,289	11,820

Table 2.13 Range of VoLL for SME users [188]

	Summer				Winter			
	Off-peak		Peak		Off-peak		Peak	
	Weekend	Weekday	Weekday	Weekend	Weekend	Weekday	Weekday	Weekend
WTA (£/MWh)	37,944	36,887	33,358	34,195	44,149	39,213	35,488	39,863

In order to take into account variations of VoLL depending on the season and time of the day, an estimation of the ranges of cost of compensation was made, following similar steps to those described above. As mentioned before, the ratio of domestic and SME users is adopted to be 74:26. Values given in Figure 2.11 show that, regardless of the time of the day or season, the highest rate of compensation for ENS comes from the interruptions in MC2 feeder type.

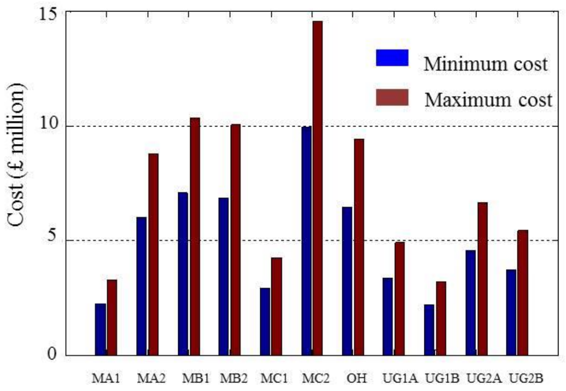


Figure 2.11 Range of cost of compensation per feeder type

2.6.1.3 Fault analysis

The statistical analysis of the number of faults can be presented with the Poisson distribution [122], since it is appropriate for describing cases in which a random event (fault in this case) happens countable number of times in a given period or area. Since the mean and variance value are the same in this case, the only parameter needed to define the distribution is the mean value (λ) of the sample set. The Poisson distribution of the number of faults per feeder for 6.6 kV and 11 kV network is shown in Figure 2.12. Comparing the intersection surfaces between the distribution of all faults and faults on 6.6 kV and 11 kV feeders, respectively, it can be seen that faults at the 11 kV level are more frequent.

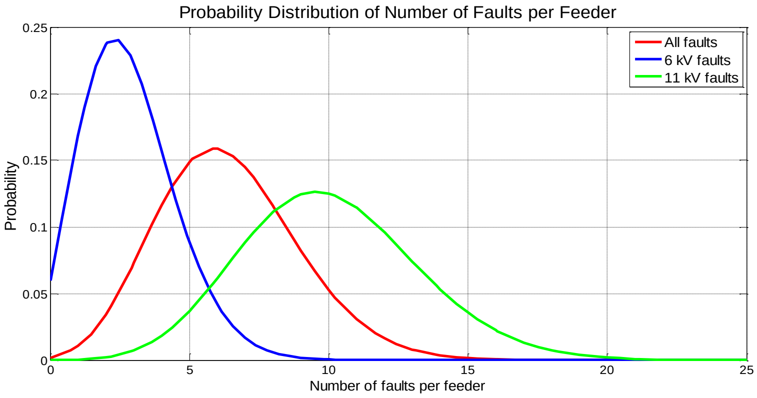


Figure 2.12 Poisson distribution of number of faults per feeder for 6.6 kV and 11 kV faults

The average number of faults per feeder in 11 kV network is around 10 (9.94), which is almost two times more than the average number of faults happening on all feeders

(6.33). On the other hand, the number of faults per feeder at 6.6 kV is around 3 (2.8), being two times less than the overall average. The reason lies in the fact that 11 kV feeders usually supply commercial and residential customers which is why they are longer, particularly in their OHL part, and therefore more vulnerable than 6.6 kV feeders. A comparison of lengths of the OHL part, underground cable and total length of 6.6 kV and 11 kV feeders is given in Figure 2.13. As shown in the figure, 6.6 kV feeders are predominantly underground cables, while 11 kV feeders are mostly overhead lines.

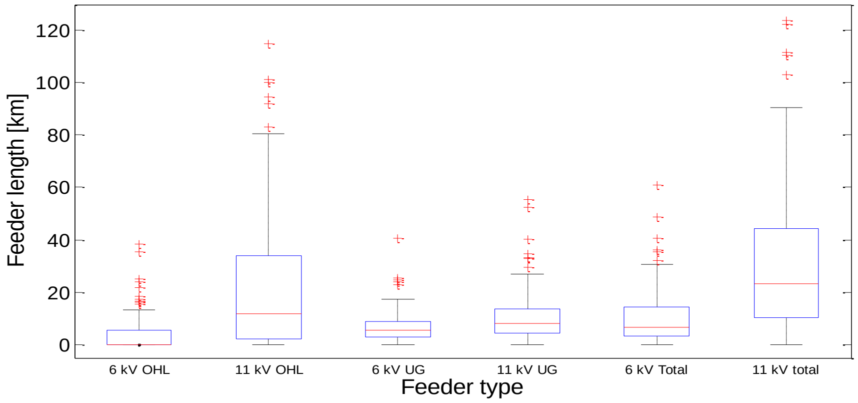


Figure 2.13 Boxplots for overhead line, underground and total length of feeders

2.6.1.4 Quality of service performance analysis

The analysis of the database continued by looking into QoS performance of the network. Looking at the QoS performance over the years 2007-2011, it can be noticed that there has not been significant change in the number of faults per feeder, as shown with the Poisson distribution of the number of faults per year in Figure 2.14. Probability distribution of the number of faults per feeder is almost the same, concentrated around the mean value of 2.6.

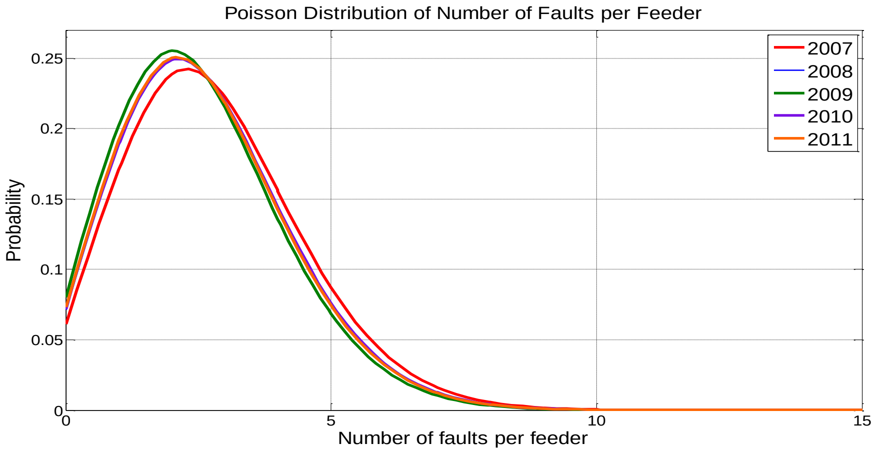


Figure 2.14 Probability distribution of number of faults per feeder over the period 2007-2011

At the same time, CDF of the average duration of interruption per feeder over the period from 2007 to 2011 is shown in Figure 2.15. Improvement in this aspect is visible through the years. For example, in 90% of the cases in 2007, faults in HV feeders caused no more than $236.6 \cdot 10^3$ minutes of interruption, while in 2011 this value decreased to $153.9 \cdot 10^3$ minutes lost, which represents improvement of around 35%.

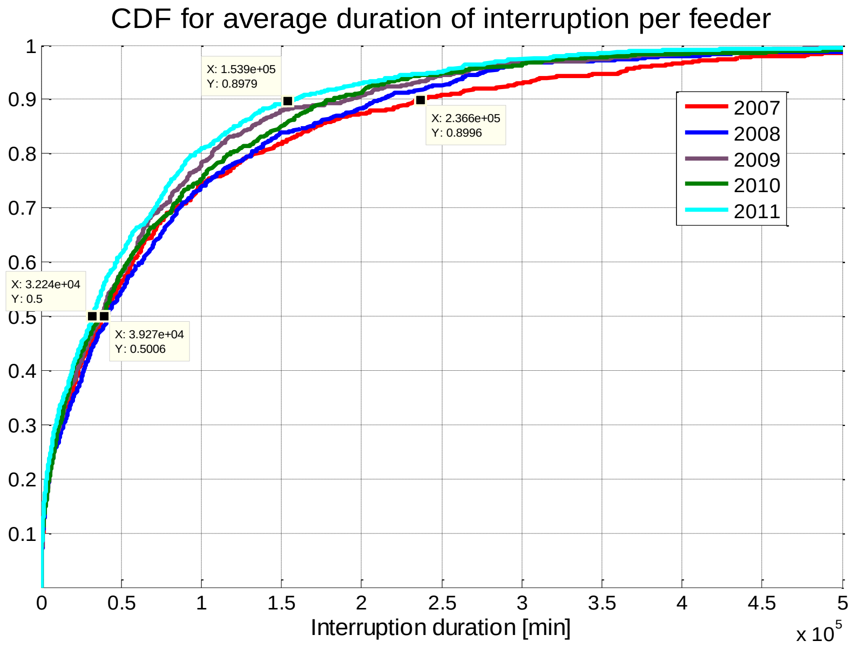


Figure 2.15 Cumulative distribution function of average interruption duration

QoS indicators were then correlated to feeder characteristics (number of customers supplied by the feeder, length (km) of the OHL part and underground part and total length (km) of the feeder). Linear regression analysis showed that correlation coefficients for the same indicators were higher for higher aggregation level. When accumulated to primary substation level, QoS indicators showed high correlation with feeder parameters, mainly with total length of the feeders and number of customers supplied from the substation. Correlation coefficients for no aggregation and different levels of aggregation are given in Table 2.14.

In order to justify high correlation between the number of faults and feeder parameters, especially at higher aggregation level, an example of learning based on available data is given with regression tree model in Weka tool, called M5 pruned (M5P) model tree. In this regression tree, the number of faults accumulated per feeder is assessed based on the length of HV feeders and the number of connected customers. Model trees are a type of regression trees with linear regression models at their leaves [118]. Regression trees are used for describing continuous attributes, unlike the classification trees which are used to describe categorical ones [178]. The result is illustrated in Figure 2.16. The first number in each bracket represents the number of instances reaching the leaf and the second number represents the percentage of misclassified instances. Each leaf contains different linear regression model, as given in Table 2.15. Num_Cust refers to the number of customers supplied by the feeder/primary substation (depending on the aggregation level) and OHL_Length presents the total length (km) of the OHL part of the feeder/all feeders connected to the substation. Similarly, UG_Length presents the total underground part length and Tot_Length presents the total length of the feeder/all feeders connected to the substation.

Table 2.14 Correlation coefficients

QoS indicator	Correlation coefficient (r) to feeder characteristics		
	No aggregation	Aggregation per feeder level	Aggregation per primary substation level
Number of faults	/	0.78	0.84
Cumulative number of interruptions during faults	0.36	0.60	0.74
Cumulative duration of interruptions (in minutes) during faults	0.26	0.60	0.72

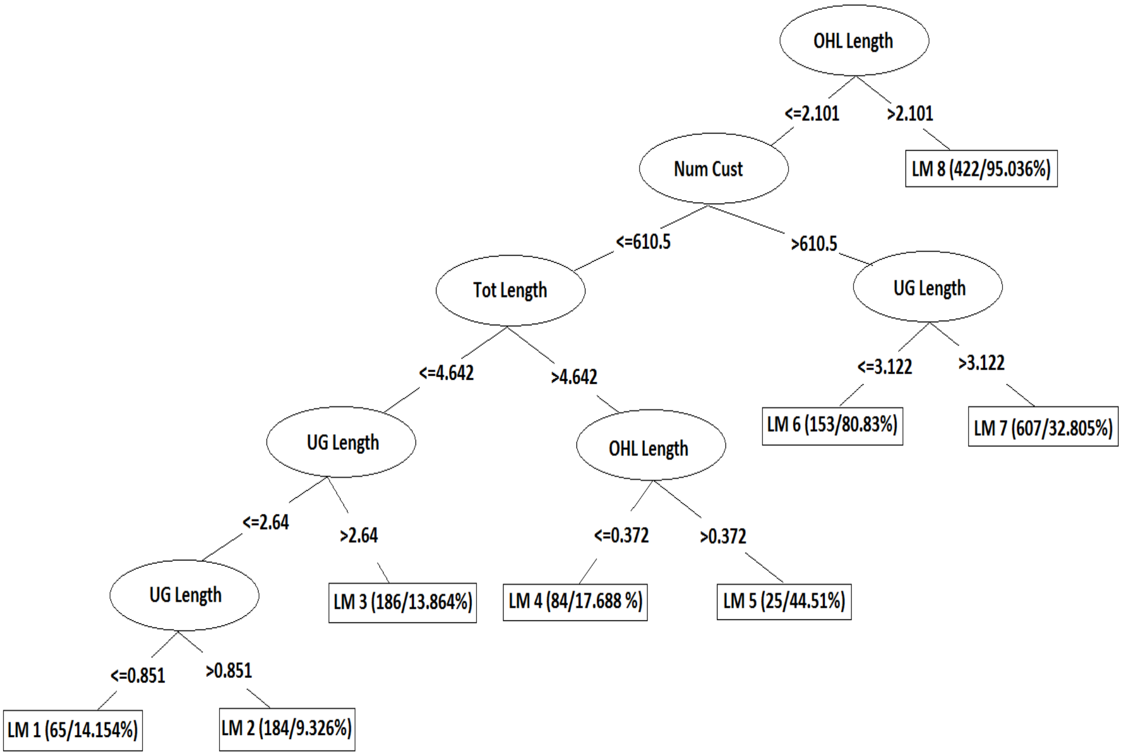


Figure 2.16 Regression tree for the number of faults per feeder

Table 2.15 Regression rules given in 8 nodes of the decision tree

LM 1	LM 2	LM 3	LM 4
$Faults = 0.0016 * Num\_Cust + 0.17 * OHL\_Length - 0.0783 * UG\_Length + 0.0643 * Tot\_Length + 1.5538$	$Faults = 0.0001 * Num\_Cust + 0.17 * OHL\_Length - 0.03 * UG\_Length + 0.0339 * Tot\_Length + 1.4589$	$Faults = 1.6734 * OHL\_Length + 0.0026 * UG\_Length + 0.0147 * Tot\_Length + 1.7958$	$Faults = 0.5582 * OHL\_Length + 0.0026 * UG\_Length + 0.3398 * Tot\_Length + 0.0115$
LM 5	LM 6	LM 7	LM 8
$Faults = -0.0061 * Num\_Cust + 0.9965 * OHL\_Length + 0.0026 * UG\_Length + 0.1359 * Tot\_Length + 5.715$	$Faults = 0.0001 * Num\_Cust + 0.2735 * OHL\_Length - 0.0051 * UG\_Length - 1.5281 * Tot\_Length + 6.1707$	$Faults = 0.0005 * Num\_Cust + 2.0667 * OHL\_Length + 0.0005 * UG\_Length + 0.2467 * Tot\_Length + 0.3354$	$Faults = 0.0009 * Num\_Cust + 0.2051 * OHL\_Length + 0.2448 * UG\_Length + 0.2031 * Tot\_Length + 4.8651$

Validation of the model was done using 10-fold cross validation, which means the data set is divided into 10 groups, so that in each of the 10 consecutive training cycles, different part of the sample (representing 10% of the overall sample) is used for testing (validation), while the remaining 90% is used for training of the tree. Finally, the averaged performance of all the 10 cycles (folds) is used for model assessment. Correlation between the training data of the M5P model is quite high ($r = 0.81$) in the case of aggregation to feeder level, with relative absolute error (RAE) and root relative squared error (RRSE) equal to 47.13% and 58.51%, respectively. The RAE and RRSE are calculated based on the predicted (P) and target (actual) values (T) of the training data, using the following expressions:

$$RAE = \frac{\sum_{i=1}^{n} |P_i - T_i|}{\sum_{i=1}^{n} |T_i - \bar{T}|} \tag{2.26}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - T_i)^2}{\sum_{i=1}^{n} (T_i - \bar{T})^2}}, \tag{2.27}$$

where P_i is the predicted value, and T_i is the target value of the i-th instance, while $\bar{T}$ is the average target value. RAE and RRSE values are relative to the variation of the target values, therefore, if their value was 100%, that would mean that the decision tree is predicting the average of the target value. Smaller error values are desirable, however in this case they are only used to compare the performance of the decision tree when it is trained with different datasets, as it will be shown in the remaining of this subsection.

The dataset for training of the tree had 1726 instances. As mentioned before, there are 8 regression rules (LM1-LM8) represented by leaves at every node of the tree, given in Table 2.15.

The same analysis was done for the number of faults aggregated at primary substation level, in which case 6 rules were generated during the training process (Figure 2.17), similarly to those in Table 2.15. Aggregation reduced the size of the training dataset, resulting in 350 instances. Correlation coefficient is higher than in the previous case ($r = 0.86$) and the relative errors are lower: 40% and 50.78% for relative absolute error and root relative squared error, respectively. This implies that a more accurate prediction of the number of faults based on feeder length and number of customers can be obtained using data aggregated per primary substation level. It can be concluded that even though the aggregation reduces the number of instances, i.e., the size of the training dataset, it improves the training of the classification tool. Since the classifier showed low correlation with the number of customers, this attribute did not participate in the tree formation.

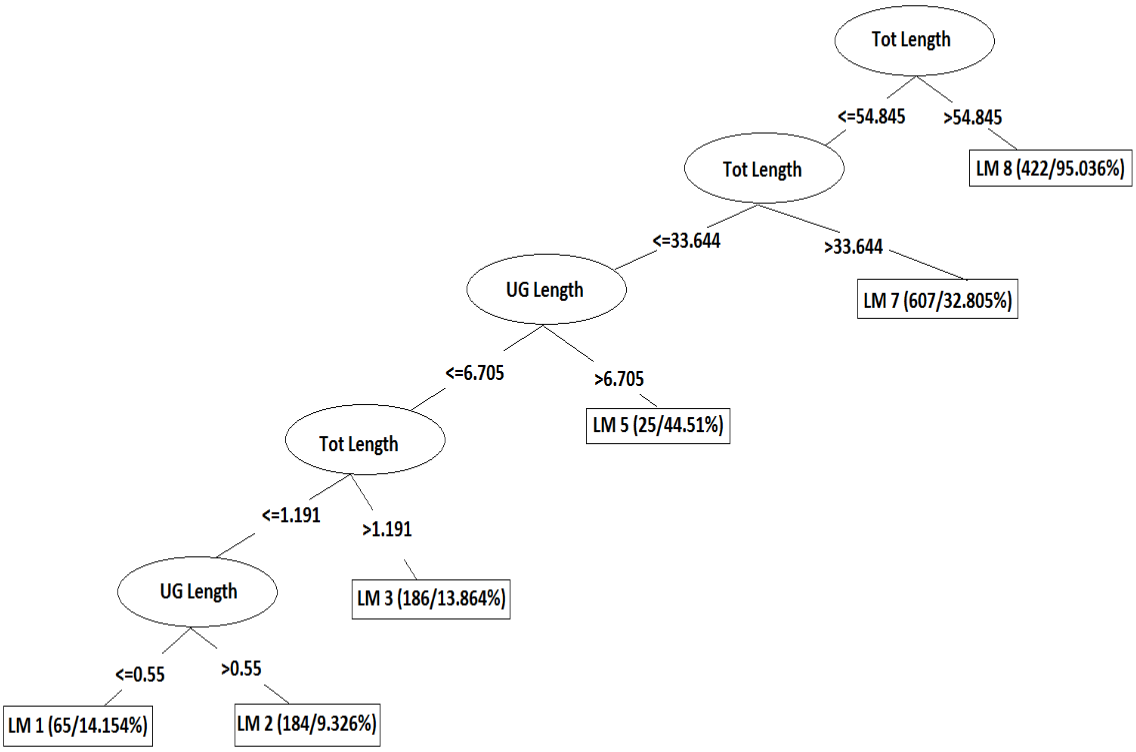


Figure 2.17 Regression tree for number of faults per primary substation

2.6.1.5 District analysis of quality of service

QoS indicators are next analysed for each of the 7 network districts, all operated by the same DSO, and shown in Figure 2.18. This was done in order to investigate possible connection between network performance and some specific characteristics of individual districts. All values are normalized according to the base, i.e., maximum values per area, as given in Table 2.16.

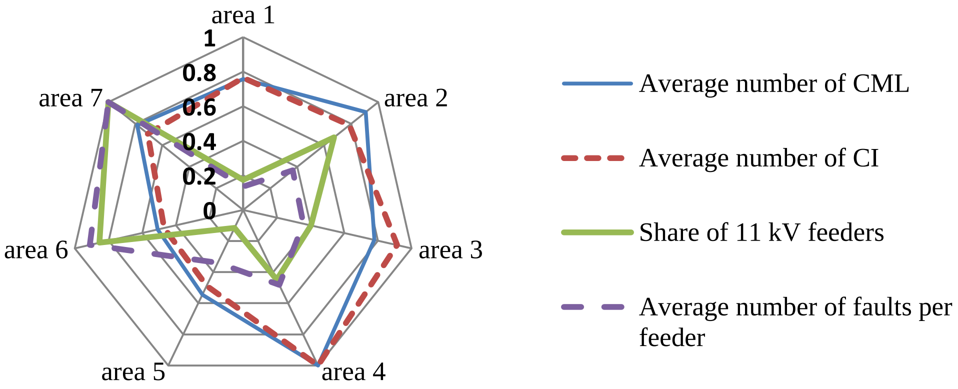


Figure 2.18 District analysis of key QoS indicators

Table 2.16 Base values for normalization in Figure 2.18

Measure	Base value
Average number of CML	94.37
Average number of CI	125.76
Share of 11 kV feeders	100 %
Average number of faults per feeder	14.02

As seen from the diagram in Figure 2.18, the number of faults per feeder is highly correlated with the share of 11 kV feeders in the HV network. The analysis of fault causes performed in Weka showed that the faults in HV distribution network were mostly caused by the asset deterioration due to the ageing, regardless of the district. On the other hand, the average number of CML and CI are mutually correlated, but not very dependent on the share of 11 kV feeders. Within the districts with poor QoS performance (areas 2, 3, 4, 6 and 7), mostly affected feeder classes are MB2, MC2 and OH. These feeders are distinguished by large total length (more than 11 km) and the share of OHL part bigger than 20%, which makes them more vulnerable to faults.

Shares of domestic and non-domestic users per district, based on the shares of 11 kV and 6.6 kV feeders, respectively (Figure 2.19), were compared with the estimated ENS in the domestic sector for all the observed districts during the five-year period (Figure 2.20). The calculation was done as follows:

1) First, number of CML was calculated for every fault and multiplied with the number of interrupted customers.

2) Value calculated in 1 was then multiplied by the average domestic consumption, standing for 1.1 kW, giving the energy lost per fault. This value was summarised for all the faults happening in each of the districts, giving the accumulated ENS for all the customers per area in the observed period (2007-2011).

3) The accumulated energy lost was finally multiplied by the VoLL for domestic sector (16.94 £/kWh), giving the cost of compensation for all the districts in the given five-year period.

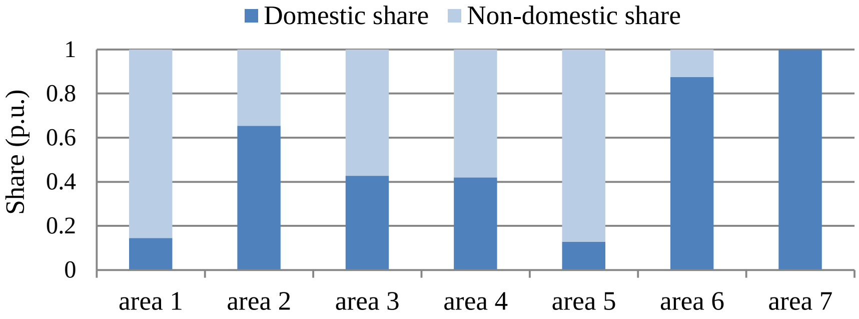


Figure 2.19 Shares of domestic and non-domestic end-users per district

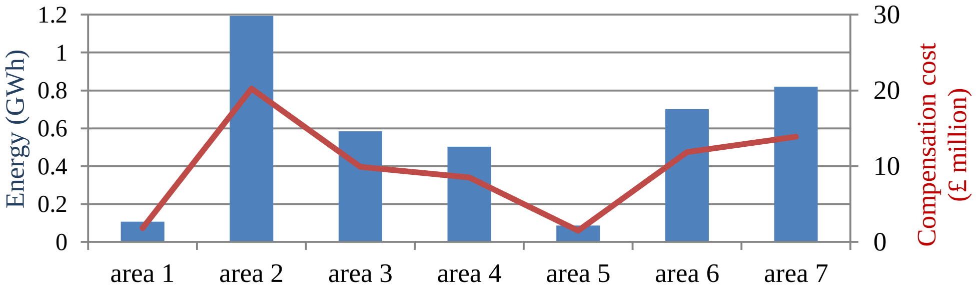


Figure 2.20 District analysis of estimated ENS and compensation cost in the domestic sector

The highest estimated expenditure for the compensation for the ENS was in area 2 due to supply interruptions, and calculated to be around 3% of the five-year profit of the DSO. This area, with domestic share of more than 60%, also showed the highest rate of ENS in domestic district, with around one million kWh of ENS. The network in this area mainly consists of MA2 and MB2 feeders, which belong to medium length feeders with less than 50% of OHL part and with deterioration due to the ageing as the main cause of fault occurrence. Distribution network in areas 3, 6 and 7 had significant total combined compensation costs (almost two times higher than the cost in area 2). It consists of OH and MC2 feeders which showed the highest fault tendency, especially since weather (wind and gale) was stated as one of the main causes of faults in these three areas.

Similarly to the analysis done according to feeder type, the range of estimated cost of compensation was done taking into account variations of VoLL depending on the season and time of the day (Table 2.12 and Table 2.13). The ratio of domestic and SME users is, as before, adopted to be 74:26. Results for the period 2007-2011 are given in Figure 2.21. As seen in the figure, estimated costs are around £2 million for districts with lower rates of ENS, between £7 million and £15 million for districts with medium rate of ENS, and from around £15 million pounds to around £24 million pounds for the district with the highest rate of ENS.

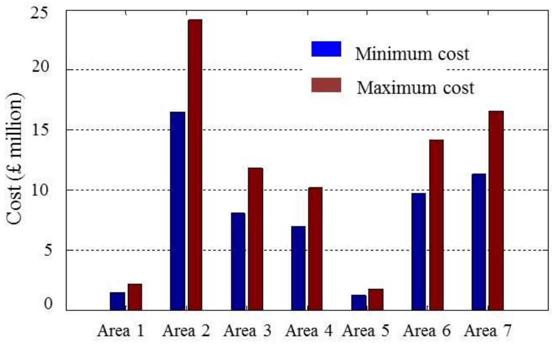


Figure 2.21 Range of revenues for domestic and SME users per district

When the ENS values over the most critical districts (2, 3, 6 and 7) get disaggregated down to a year level, as in Figure 2.22, it can be seen that the overall excessive ENS in district 2 is dominated by very large ENS in one year. Therefore, further analysis should be performed to investigate possible reasons for this.

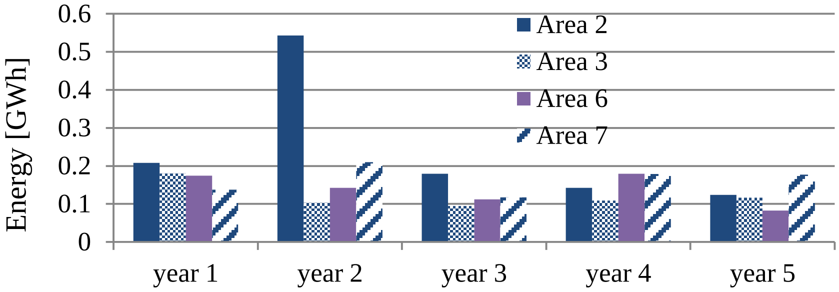


Figure 2.22 Amount of ENS during five-year period in some districts

2.6.2 LV analysis

As part of the LV analysis, probability distribution of the number of faults per feeder in LV network was compared to the distribution of faults causing damages on feeders and faults which do not cause them. As seen in Figure 2.23, all three distributions have close to normal distribution shape, with different mean values. The sizes of intersection areas between distributions show that the majority of faults on LV feeders do not cause damage.

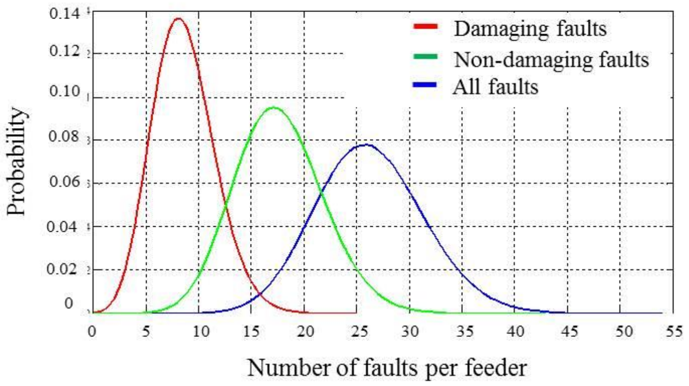


Figure 2.23 Probability distribution of faults in LV feeders

From the CDF of the number of faults per feeder (Figure 2.24), it can be seen that in 90% of the cases the number of faults causing damages is less than 15, while the number of non-damaging faults is higher, about 27 for the majority of cases. The analysis of fault causes in Weka showed that the majority of damaging faults were caused by the asset deterioration due to the ageing, corrosion or third party. Similarly to the fault statistics per year in HV distribution network with somewhat constant rate, the number of faults in LV network also did not change much over the five-year period.

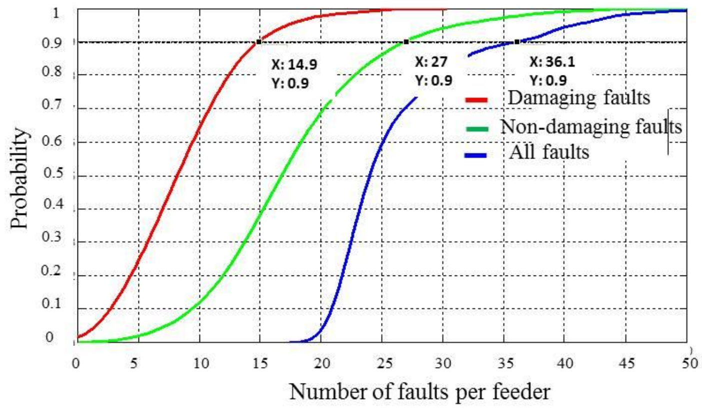


Figure 2.24 Cumulative distribution function of the number of faults per feeder

Histograms in Figure 2.25 and Figure 2.26 refer to the cumulative duration of interruptions per fault and the number of interruptions per fault, respectively, classified into faults that caused damage on LV feeders, and those that did not.

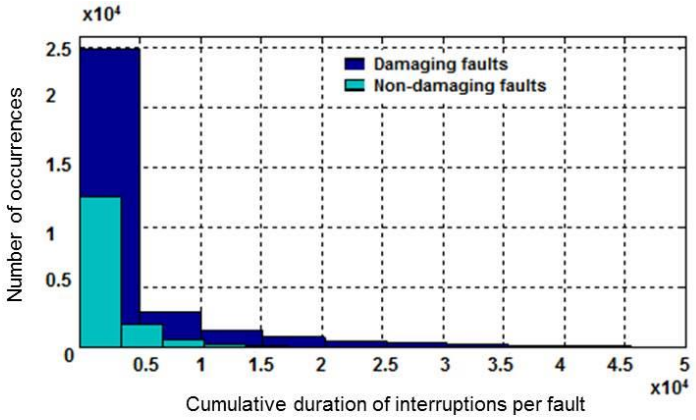


Figure 2.25 Histogram of the cumulative duration of interruptions during damaging and non-damaging faults

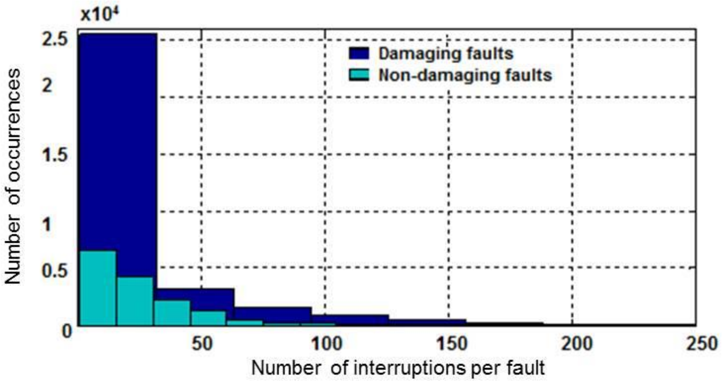


Figure 2.26 Histogram of number of interruptions during damaging and non-damaging faults

As seen in Figure 2.25, in case of interruptions with cumulative duration of up to 3.5 days (5,000 minutes), the frequency of damaging faults is around two times larger than the frequency of non-damaging ones. If the cumulative interruption lasts between 3.5 and 14 days (20,000 minutes), it is also more probable that the fault caused damage. The figure also shows that cumulative duration of interruptions longer than 14 days almost always happens due to damaging faults. Similarly, as illustrated in Figure 2.26, among faults causing less than 40 customer interruptions (i.e., 40 customers interrupted), the frequency of damaging faults is drastically bigger than the frequency of non-damaging faults. The number of faults that caused more than 100 interruptions is much smaller and mainly caused damages on feeders.

The radar diagram of LV network performance among different districts is shown in Figure 2.27. As seen from the diagram, area 1 shows the worst performance with respect to the most of the indicators, together with areas 2 and 3 that show a high rate of CML and CI on both, the average and cumulative level. The typical percentage of faults that cause damages on feeders, aggregated at primary substation level, is between 80 and 100%.

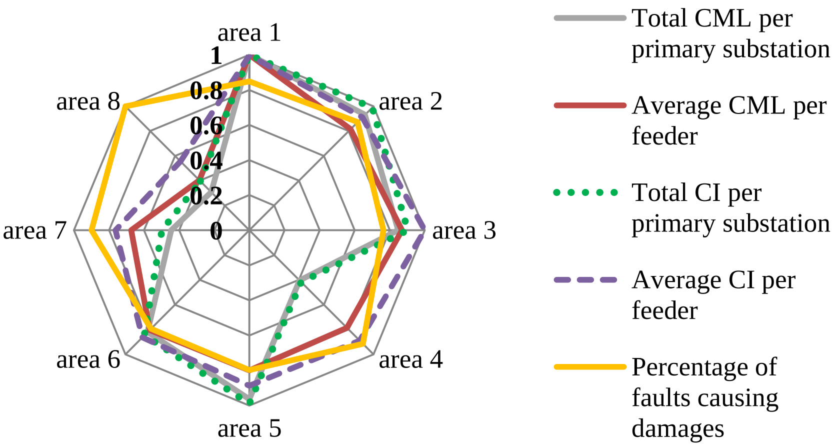


Figure 2.27 District analysis of key QoS indicators

2.6.3 Discussion

As the database considered in this case study is not particularly large, only some of the data analytics methods described in Section 2.2 were applied, namely linear regression and decision trees. Results of the correlation analysis presented in this section have shown that QoS parameters very much depend on feeder characteristics. Therefore, prediction methods, such as decision trees or linear regression, could be used to form a model for QoS indicators estimation based on some asset characteristics. At this point, this type of model would show significant errors due to a relatively small number of instances in the training data set. That means that prediction models bring more benefit in case of effectively larger sample size, in this case larger number of feeders observed or larger historical data.

As the analysis showed on the example of a typical database owned by a DSO, useful information about the network performance, both spatial and temporal, can be obtained from raw data. This is done by appropriate level of aggregation of data (by feeder or by primary substation in this case) and classification (by feeder class, for example). The first condition that has to be fulfilled is that the data is “cleaned” from instances with outliers which could drastically change the output of the data analysis.

It can be concluded that, with the use of data mining methods, estimation of future network performance can be facilitated based only on some static data, such as feeders’ attributes. This can be of great help to asset managers in terms of decision making and significant savings for the utility. As the results showed, investments can be focused on a particular type of asset (in this example, a feeder class) or a network district showing lower QoS performance indicators. Statistical analysis can also show some interdependencies of events, as shown with histograms of failures and damages on LV feeders.

2.7 Summary

This chapter gave an overview of data mining methods typically used in power system studies, mainly focusing on distribution network. An analysis was made of the presently collected data and data types that may find their use in the future analysis and operation of the distribution system. Some of the most important tasks when dealing with the ever-growing databases in power utilities are to determine the key data types (data prioritisation), their optimal sampling step, the frequency of their collection and the appropriate aggregation level. This is a necessary step towards obtaining a trade-off between the usefulness ("informativeness") of data on one side, and the size (and cost) of the databases and communication lines required to accommodate these on the other. An overview and critical appraisal of different data analytics methods, including text mining, for application in distribution system studies represent the first original contribution of this thesis.

An overview of data provided by SMs was made next, based on their reported technical specifications and requirements enforced by the regulatory agencies. The analysis of future DSO's requirements and SM specifications shows that SMs play an important role in the development of the smart grid. As the SM rollout will have different success in different distribution networks, methodologies should be developed to obtain as much information as possible from residential areas even with limited SM coverage (i.e., observability) of the end-users. Identification of data needs in future distribution networks and the extent to which smart metering can help with meeting these needs represent the second original contribution of this thesis. Data mining methods could also reveal groups of data that are more relevant to extraction of specific information. For example, in the presented case study used to illustrate the application of data analytics methods in distribution network analysis, regression tree analysis revealed that the number of faults at the substation level does not have a noticeable correlation with the total number of customers supplied by the substation. Therefore, this type of data does not have to be collected for the purpose of assessing the expected number of faults at a substation based on feeder characteristics. In addition, data mining can support decision making in asset management and enhance savings for the power utility. The study showed that critical assets, as the “candidates” for monitoring or replacement, can be identified based on their characteristics and using regression models built on historical data.

3 Advanced Demand Profiling

3.1 Introduction

The rollout of smart meters (SMs) in distribution networks should enhance the observability of the demand side. In order to make this observability useful to the distribution network operator (DNO) and/or other demand response (DR) responsible parties, information about time varying demand composition and its flexibility (both in close to real time and forecast) should also be provided. The missing piece of information necessary for estimating or predicting the demand side flexibility can be obtained by more detailed monitoring of the end users, e.g., via non-intrusive load monitoring (NILM) methods or by enabling communication between SMs and smart home devices. Deployment of these technologies is still at its infancy (if individual pilot sites are neglected) and requires additional investments by the distribution network utilities and the end-users. The advanced demand profiling should enable confident assessment of demand composition and its flexibility at the aggregation (e.g., substation) level, including both monitored and non-monitored end-users. Therefore, the main question discussed in this chapter is how many, i.e., what portion of, end-users should be monitored in detail to allow for advanced demand profiling. In other words, how many users in an aggregation would have to provide close to real time appliance-level consumption data in order to estimate/forecast the composition of aggregated demand? The answer to this question is provided in this chapter by developing artificial neural networks (ANN) based methodology for aggregated demand decomposition.

Decomposition of the aggregated demand provides information about the contribution of different load types (induction motors, lighting, resistive loads, etc.) to the total load demand, and hence the flexibility (controllability) of the demand. The two main uncertainties associated with load decomposition relate to customers' behaviour and quality and availability of SM data (including missing samples and noise caused by monitor faults or communication problems). All these are aggravating factors affecting the accuracy of load decomposition. As the desired level of the accuracy of the result depends on its application, the analysis given in this chapter aims at discovering to what extent different factors influence the accuracy of demand decomposition. Even though the proposed methodology observes residential demand only, it is equally applicable to industrial, commercial or mixed demand sectors.

3.2 Demand decomposition

The output of the demand decomposition process, performed at the aggregation or substation level, provides the information about the time-varying load shares (in per unit or percentage) of different load categories within the (time-varying) total active or reactive demand. Following methodology discussed in [86], load categories in this study are defined as groups of appliances with similar voltage-dependent steady-state and dynamic load characteristics. Furthermore, load categories are divided into controllable and uncontrollable, based on their potential to be shifted in time. The controllability of some loads is disputable, as in the case of lighting loads - although, generally, they are considered to be uncontrollable, some of them can be dimmable and therefore controllable. Thus, the given classification should be taken as illustrative only, as it could vary to a certain extent for different applications. This study considers as controllable all the appliances that may be a part of direct load control (e.g., fridges, water heaters) or incentive-based DR programs (e.g., washing/drying machines), i.e., appliances that can be controlled/shifted automatically or by the users.

According to the most commonly used appliances in residential sector in the UK, six categories are recognized in this methodology and presented in Table 3.1. They include single-phase constant torque induction motors (CTIM1), single-phase quadratic torque induction motors (QTIM1), controllable resistive loads (R_C), uncontrollable resistive loads (R_{UC}), switch-mode power supply (SMPS) loads and Lighting. The full list of appliances, apart from heating, ventilation and air conditioning (HVAC) units, was adopted from the CREST residential load model [111]. The same model was used to generate individual daily load profiles of the end users, which served as a realistic representation of data streams coming from the SMs. Controllable loads mainly consist of thermostatically controlled loads which do not affect customers' comfort drastically [190] (space and water heaters, fridges, freezers), and wet appliances (washing machines).

Table 3.1 Load categories and corresponding types of domestic appliances

Load controllability	Load categories	Residential appliances
Controllable	1. CTIM1	HVAC, dish washer, tumble dryer, washing machine, washer-dryer, vacuum cleaner
	2. QTIM1	Chest freezer, fridge-freezer, fridge, upright freezer
	3. R_C	Water heater, electrical shower, storage heater
Uncontrollable	4. R_{UC}	Iron, hob, oven
	5. SMPS	Answer machine, CD player, Clock, telephone, high fidelity (HiFi) appliances, Fax machine, PC, printer, TV, VCR-DVD, receiver, microwave
	6. Lighting	Lighting

The diagram in Figure 3.1 presents the main steps of the methodology for demand decomposition in a smart metering system with partial coverage, i.e., where only some users have per-appliance monitoring, as a fairly realistic scenario in the future distribution grid. As mentioned earlier, per-appliance monitoring can be achieved either by a NILM method, or via communication between smart home appliances and SMs. Two assumptions are made in this respect:

i) SMs can record the real power of individual appliances only, while reactive power is derived probabilistically, as it will be detailed in Section 3.4;

ii) Forecast of the total consumption (real and reactive power) at the aggregation level, i.e., at the substation (block {5} in Figure 3.1), is already available.

As an initial step before the demand decomposition process, the SM data is pre-processed and aggregated at the data concentrator point (block {1}). Following the first assumption, the part of the consumption which has sub-metering can be decomposed into categories or controllable/uncontrollable load by simply aggregating consumption of appliances belonging to the same category (block {2}), as detailed in Table 3.1. It should be noted that a total of 1000 households/end-users supplied from the substation, including those with and without sub-metering, is used to illustrate the approach.

At the next step, the ANN is trained with the available sub-metering data in order to be able to “recognize” the load composition based only on aggregated active and reactive load curve of the monitored users. Once trained, the ANN (block {4}) uses forecast of the total active and reactive load at the bulk point (block {5}) as the input, and gives corresponding load composition, i.e., weighted factors of each load category, as the output (block {6}). Figure 3.2 illustrates the decomposed daily loading curve (DDLC) for an aggregation of 1000 users. The main steps of the methodology are discussed in the following sections.

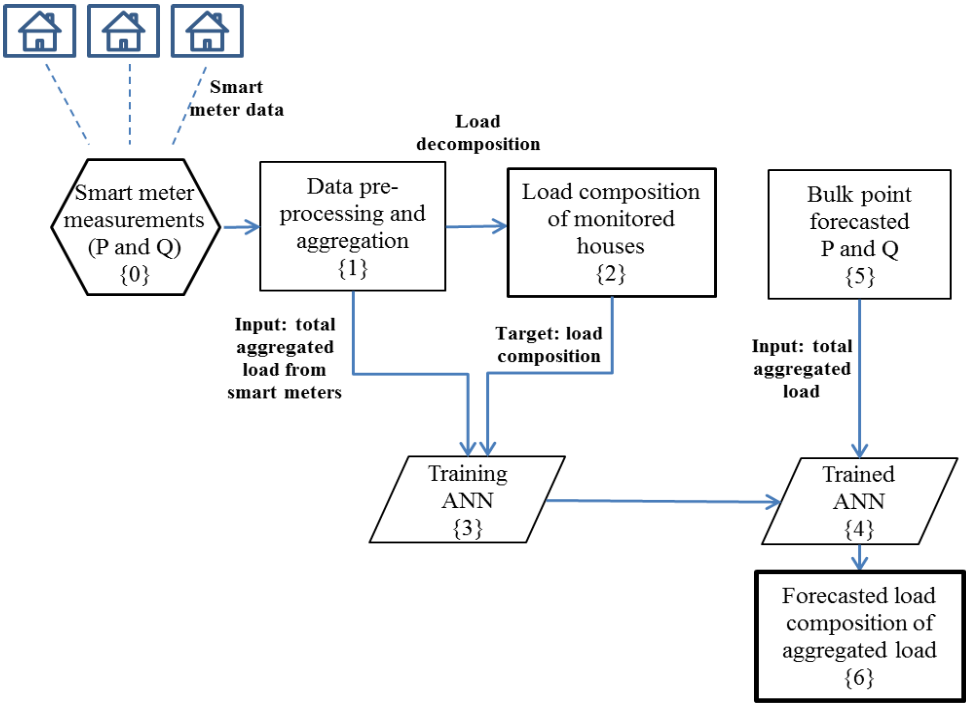


Figure 3.1 Flow chart for load disaggregation in case of smart metering system with partial coverage

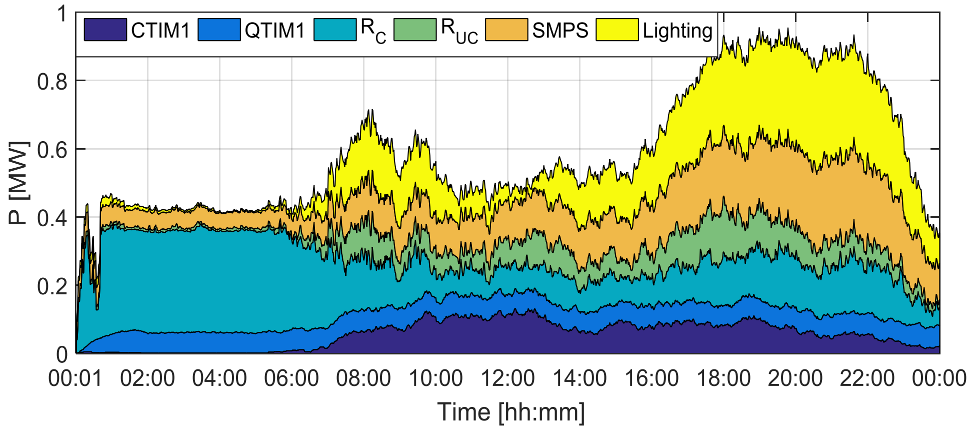


Figure 3.2 Aggregated smart meter data in an aggregation with full SM coverage (1000 houses)

3.3 Data pre-processing

In order to present as realistically as possible the future smart metering system, two assumptions are made:

i) *There are missing samples in the data streams coming from SMs*

According to [191], up to 20% of active load measurements at substation points are inaccurate. Therefore, it is assumed that there is 20% of missing data in the overall data coming from SMs, due to either sensor faults or communication problems. Missing data are presented as missing “chunks” of different lengths, distributed over the data streams in a random manner, and respecting the constraint of the 20% of the total missing data.

ii) *Different SMs have different sampling steps (with one, ten, thirty or sixty minute granularity)*

This assumption is based on [192] where it was reported that the active and reactive consumption could be measured over periods from 1 to 60 minutes. Following this, 1000 SMs in the aggregation are randomly assigned to one of these 4 groups.

One minute is taken as the reference sampling rate, as it avoids under-estimation of electrical consumption and provides sufficient data for detailed modelling of distribution networks [111]. Based on the two aforementioned assumptions, the missing samples in data streams are the consequence of both actual missing data, and the different sampling steps of different SMs. For instance, in case of 10 minute-based sampling, there are 9 samples (minutes) missing per every sampling step. As the proposed methodology uses aggregated SM data, the first step is to pre-process “raw” data streams, i.e., restore the missing data and adjust the granularity of all the streams to minute-based samples. In the preliminary studies, the noise had been added to some data streams using Gaussian White noise with relatively low signal-to-noise ratio. The noise was then filtered using locally weighted polynomial regression method [193]. This however, had negligible effect on the accuracy of the results and hence the noise was not taken into account in further studies, including these.

For comparison purposes, missing data samples, resulting from both faults (data not sent or not delivered) and higher sampling steps, are restored using two different methods: linear interpolation (LI) between existing samples and weight adjusted k-nearest neighbour (WAkNN) method [122]. The performance of the data restoration methods is assessed by calculating the relative error. The relative error is calculated by comparing the total load curve restored by one of the two methods with the load curve in case of full data availability (100% SM coverage with no missing data - total load in Figure 3.2) over a one day period.

WAkNN method requires a set of training data (usually historical data) which is then used to restore missing samples in the test data using distance (e.g., Euclidean distance [118]) minimisation. If a training object has smaller distance from the test object, this training object will get a higher weight. In this study the method calculates the missing value by weighting the 5 most similar (closest ones based on the Euclidian distance) samples from the training data. It is assumed that the historical measurements of the total active load, aggregated from SMs from the last 7 days preceding the day with missing data, are available as the training data for WAkNN. Therefore, the 5 most similar samples from the historical dataset are weighted to restore the missing one. In this example, the two attributes of the recorded data are the load and the time label (ranging between 1 and 1440 for every minute in the 24 hour period). Therefore, the 5 most similar samples are those 5 samples which have the same time label as the one with the missing load value.

As an illustration, Figure 3.3 shows a DLC whose part with missing data (20% of data in total is missing in this example) is restored using either LI or WAkNN method. In this example, aggregation of 50 users is shown.

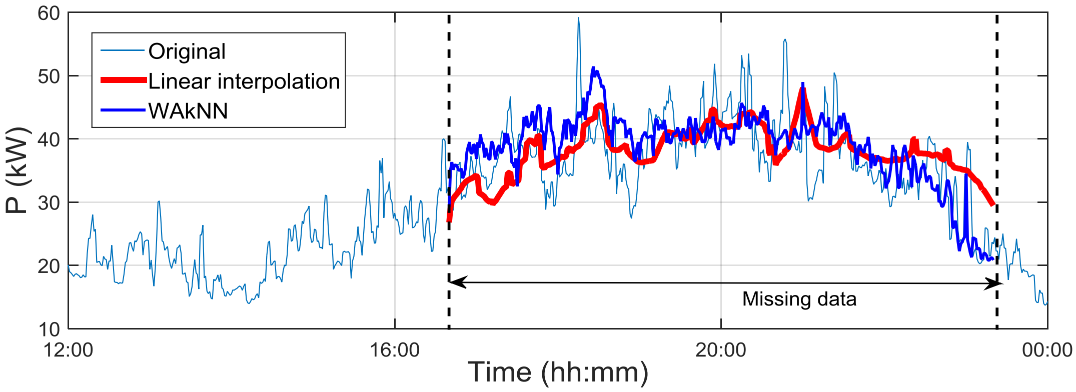


Figure 3.3 Loading curve restored using linear interpolation and WAkNN method

The variability of the DLC can be more or less pronounced, depending on the aggregation level, as illustrated in Figure 3.4. The difference in DLC between days is very visible at the lower aggregation levels of 10 or 200 SMs, while at the higher aggregation level it can only be seen between working and non-working days.

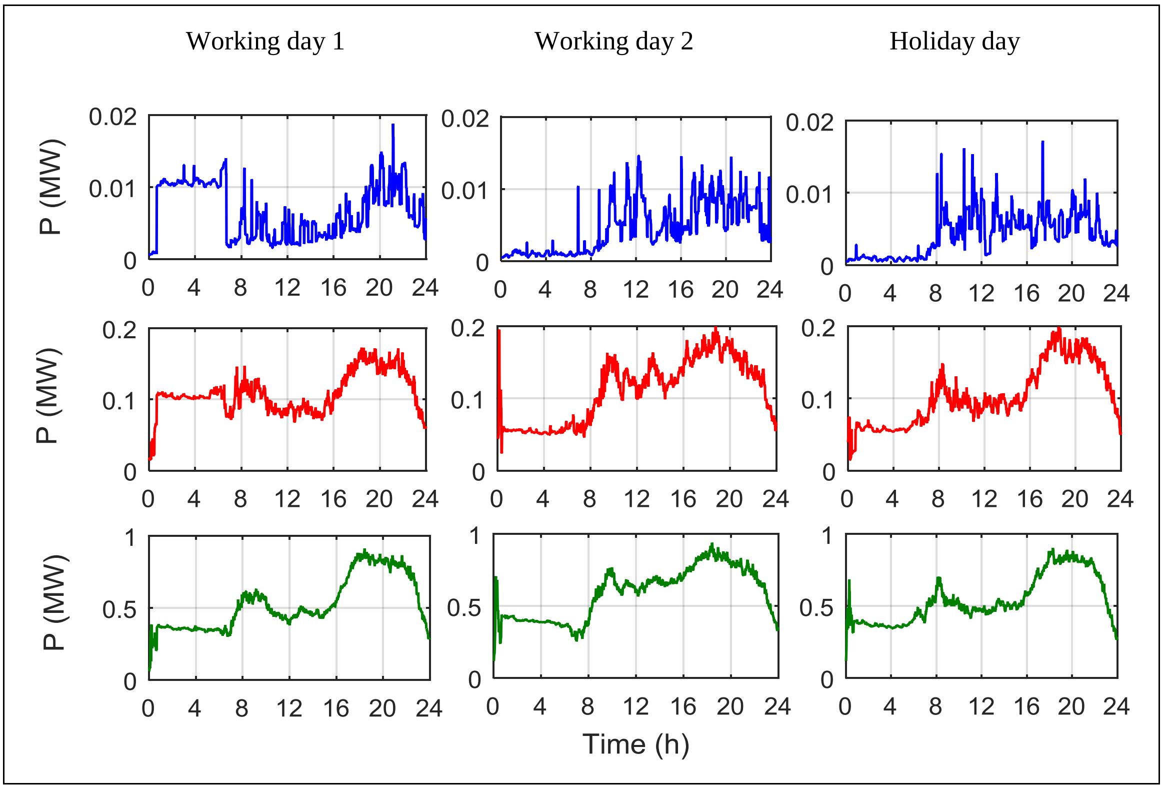


Figure 3.4 Daily load curves for aggregation of: 10 houses (first row), 200 houses (second row) and 1000 houses (third row)

Figure 3.4 also illustrates how challenging the load forecasting can be in case of lower aggregation levels, due to a large variability of load profiles from day to day, even when the sample of customers is the same. The randomness of DLC at lower aggregation level is affected by the DLC of the individual users. Figure 3.5 presents the difference between maximum and minimum load at each minute of the day, for one user observed during four consecutive Mondays in August 2015 (Pecan street dataset [24]) normalized based on the average Monday demand in the observed month. In order to evaluate this variation, variance (var) of demand is calculated during the four observed Mondays, as follows:

$$var = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - D)^2 = 0.7284 \text{ kW}^2, \qquad (3.1)$$

where n is the number of samples (here 4 days times 1440 samples), d_i is the i-th sample of the time series, and D is the average demand of the four observed Mondays. Based on the Figure 3.5 showing the relatively high variation value, it can be concluded that there is almost no repetitiveness in daily consumption even for the same user and for the same day of the week.

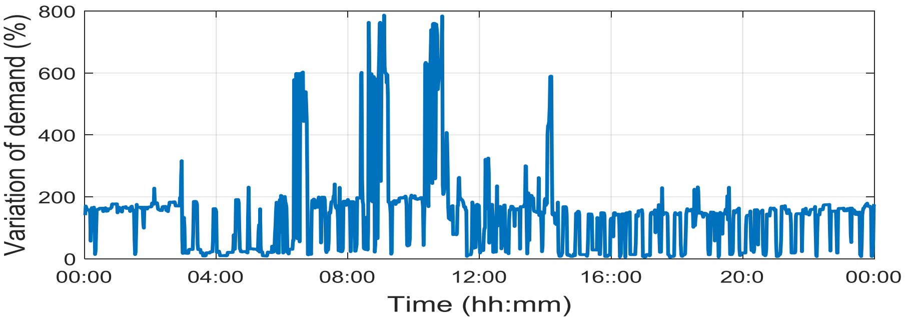
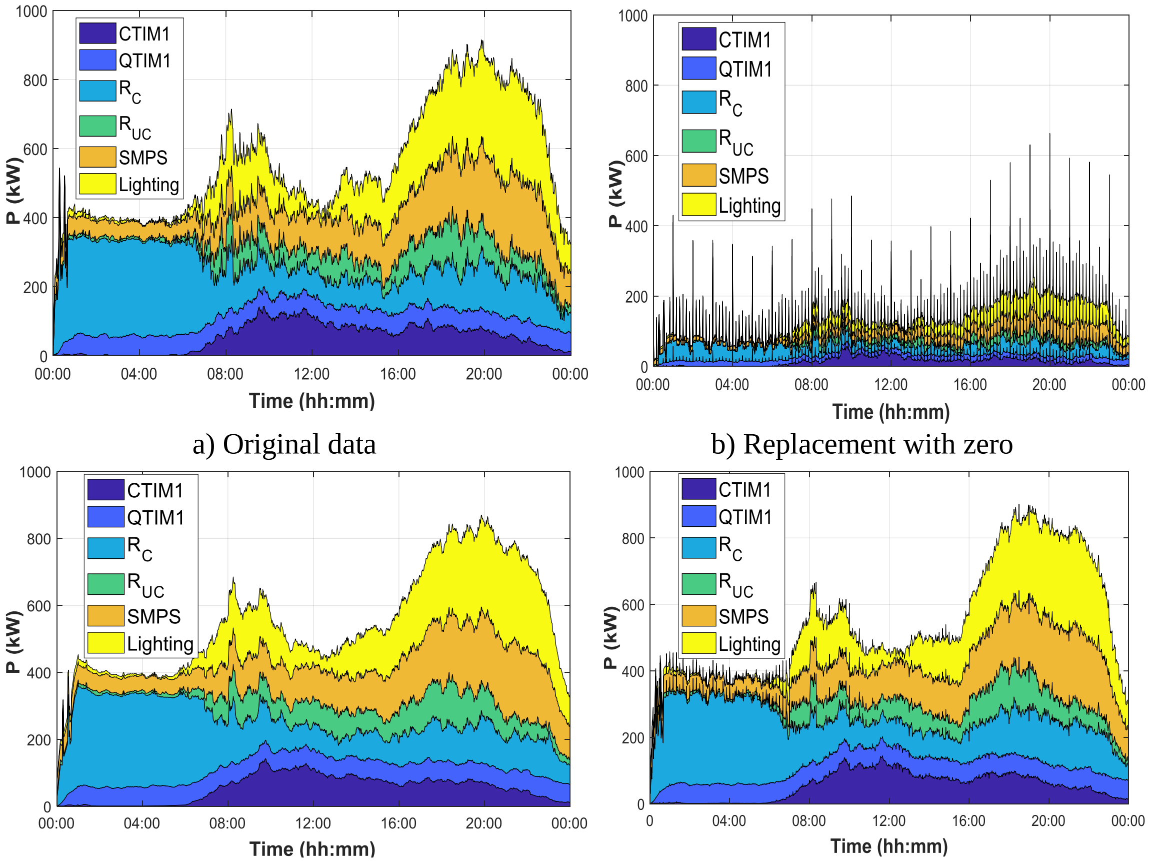


Figure 3.5 Variability of load during four Mondays of the same month of the individual residential user

In order to compare the two data pre-processing methods, Figure 3.6 presents the decomposed daily load curve (DDLC) in cases of full (original) data with no missing values, dataset with 20% missing data without restoration (missing, i.e., “NaN” values are only replaced with zero values) and data restored by the two aforementioned methods, for the aggregation of 1000 SM data. As both methods (LI and WAkNN) visually give reasonably good results, their accuracy is compared for three levels of aggregation: 1000, 200 and 50 households, and given in Table 3.2.



a) Original data

b) Replacement with zero

c) Conditioning with LI

d) Conditioning with WAkNN

Figure 3.6 Decomposed daily load curve in case of: original data (a), missing data replaced with zero (b), missing data restored by linear interpolation (c) and missing data restored by WAkNN (d)

Table 3.2 Accuracy of the two data restoration methods

Method	E_{max} (%)	E_{avg} (%)	RMSE (%)
1000 houses			
LI	**53.39**	3.97	5.44
WAkNN	35.52	**4.54**	**5.52**
200 houses			
LI	123.92	8.40	11.33
WAkNN	**298.22**	**14.87**	**14.35**
50 houses			
LI	85.63	14.33	20.66
WAkNN	**676.09**	**44.25**	**26.55**

E_{max} and E_{avg} stand for the maximum and average values of relative errors across 1440 minute-based samples (one day) and $RMSE$ stands for the root mean square error, defined as follows:

$$RMSE = \frac{\sqrt{\frac{\left(x_i - x_{0,i}\right)^2}{N}}}{\bar{x}} \qquad (3.2)$$

where x_i is the calculated (imputed) value, $x_{0,i}$is the actual value, and N is the number of samples (here, equal to 1440). $RMSE$ is normalised based on the mean daily power value of the original data set ($\bar{x}$), at the corresponding aggregation level. As seen in Table 3.2, in most cases LI method showed higher accuracy (lower accuracy in each case is highlighted in red), and, as expected, the accuracy decreased with lower aggregation level.

In addition to comparison of disaggregation into load categories, the division into controllable/uncontrollable load is also performed over the aggregation of 1000 homes. This is done to illustrate how the two data restoration methods influence the accuracy of demand decomposition into controllable/uncontrollable load. The results of demand sub-division into controllable/uncontrollable load are presented in Figure 3.7, while the corresponding relative errors are given in Figure 3.8 and Figure 3.9 for total load and controllable/uncontrollable load, respectively. Similarly, Table 3.3 shows the maximum and average relative errors, as well as the normalized RMSE (lower accuracy is highlighted using bold red font). It can be seen that the assessment of uncontrollable load is more accurate in general, while the two methods show different performance, depending on the type of the error. Nevertheless, as LI method showed higher accuracy in the restoration of total load curve, as well as the controllable load curve, it will be used for restoration of missing data in the rest of the studies given in this chapter.

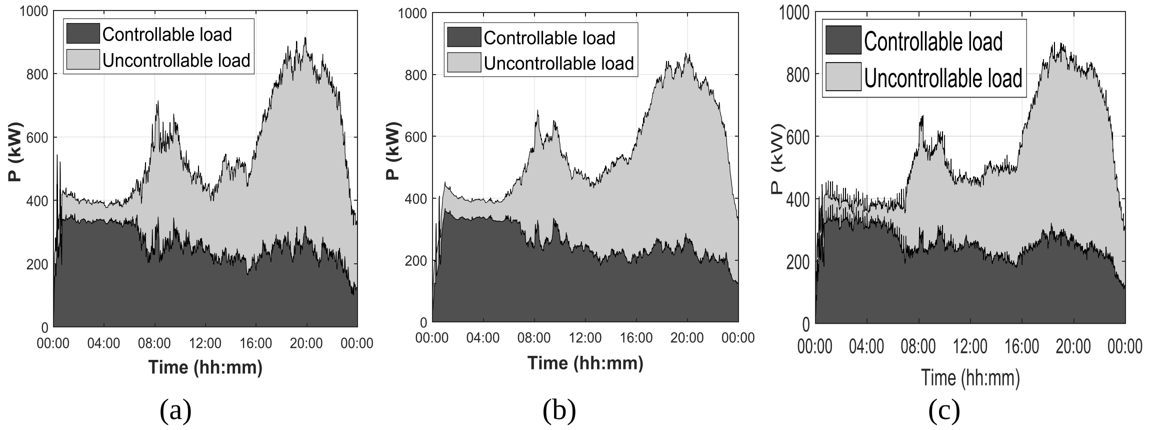


Figure 3.7 DDLCs of: (a) perfect data streams, (b) incomplete data streams conditioned using LI and (c) incomplete data streams conditioned using WAkNN

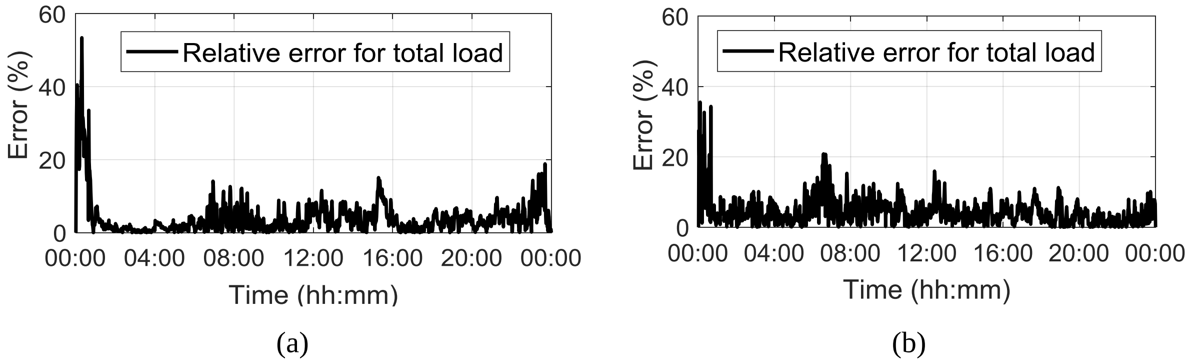


Figure 3.8 Relative errors for total load when missing data is restored using: (a) LI method and (b) WAkNN method

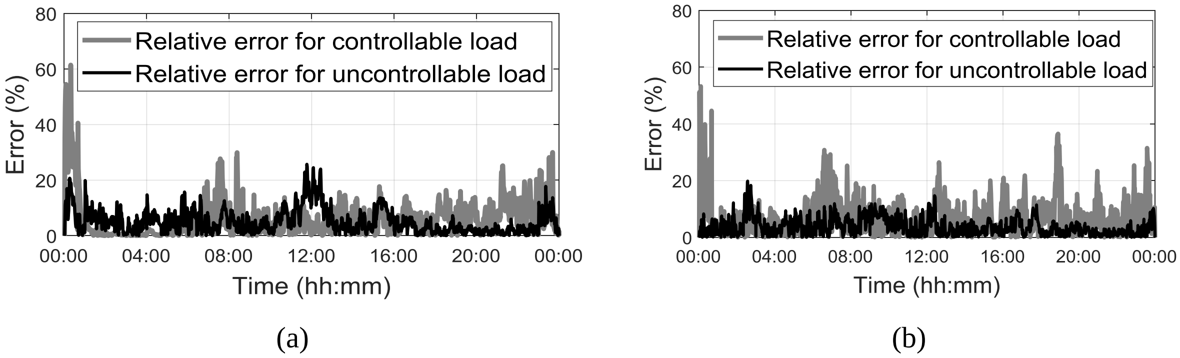


Figure 3.9 Relative errors for controllable/uncontrollable load (C/UC) when missing data is restored using: (a) LI method and (b) WAkNN method

Table 3.3 Accuracy of the restoration methods for controllable/uncontrollable load

Error	Restoration method	Controllable load	Uncontrollable load
Emax (%)	LI	**61.46**	**25.72**
	WAkNN	53.25	19.86
Eavg (%)	LI	6.15	5.15
	WAkNN	**7.15**	3.81
RMSE (%)	LI	4.32	**2.78**
	WAkNN	**4.60**	2.47

3.4 Probabilistic generation of reactive load curve

In order to make a complete profile of aggregated load in an area, both active and reactive load measurements are needed. In cases where SMs do not collect reactive power data, it can be assessed probabilistically. A bottom-up approach is therefore taken in this study, by considering the possible ranges of power factors (PFs) for different home appliances, adopted from manufacturers' websites. Assuming that the active demand of individual appliances is monitored, reactive demand of the monitored users can be derived probabilistically for each appliance and in every time step by running Monte Carlo simulations over the most common values of PF for each type of residential appliance. PF value for each appliance in each time step is sampled 100 times using randomization with uniformly distributed samples within the considered range to account for PFs of different devices and possible variability of this PF from one operating condition to the other. Then, the set of probabilistic reactive load values $Q_{k,j,i}$ is calculated for each appliance j in each time step i based on the (deterministic) active load of the appliance P_{ji} and the corresponding probabilistic values of the $PF_{k,j,i}$ ($k = 1 \div 100$ in each time step), as follows (assuming all the loads are inductive):

$$Q_{k,j,i} = P_{ji} \cdot \sqrt{1 - {PF_{k,j,i}}^2}/PF_{k,j,i} \tag{3.3}$$

The next step is to decide, at each time step, which value from the probabilistic range (namely, the mean value or the most probable value) will be adopted as the resulting one and used as the reactive demand of each appliance, load category, and consequently, the total reactive demand of the end user. In order to develop the approach, real load data from a 15 kV substation was chosen for testing, following the steps given in the diagram in Figure 3.10. The main steps in the flowchart, based on the available measurements at the pilot site, are highlighted in red, while the "background" steps, using CREST tool [111] for deriving PF values for load categories based on appliances' PF, are presented in blue.

The available substation data included half hour based measurements of voltage and current and yearly information on typical values of the PF based on the period of the day (48 samples), type of the day (working day, Saturday or Sunday) and season of the year (winter/summer). These data allowed for the calculation of active and reactive load values, as presented by *boxes 1*, *2* and *3* in the flowchart in Figure 3.10. Although the PF is not measured, but given by the electric utility, it is adopted as the correct value used for calculating reactive load. As there was no sub-metered data in the dataset, i.e., no information about the shares of load categories in total demand, the demand was decomposed using probabilistic approach (i.e., ANN trained with probabilistically generated data originating from statistical data about the electricity usage in UK domestic sector [194]). This step, described in more detail in [39], is represented by *box 4* in Figure 3.10.

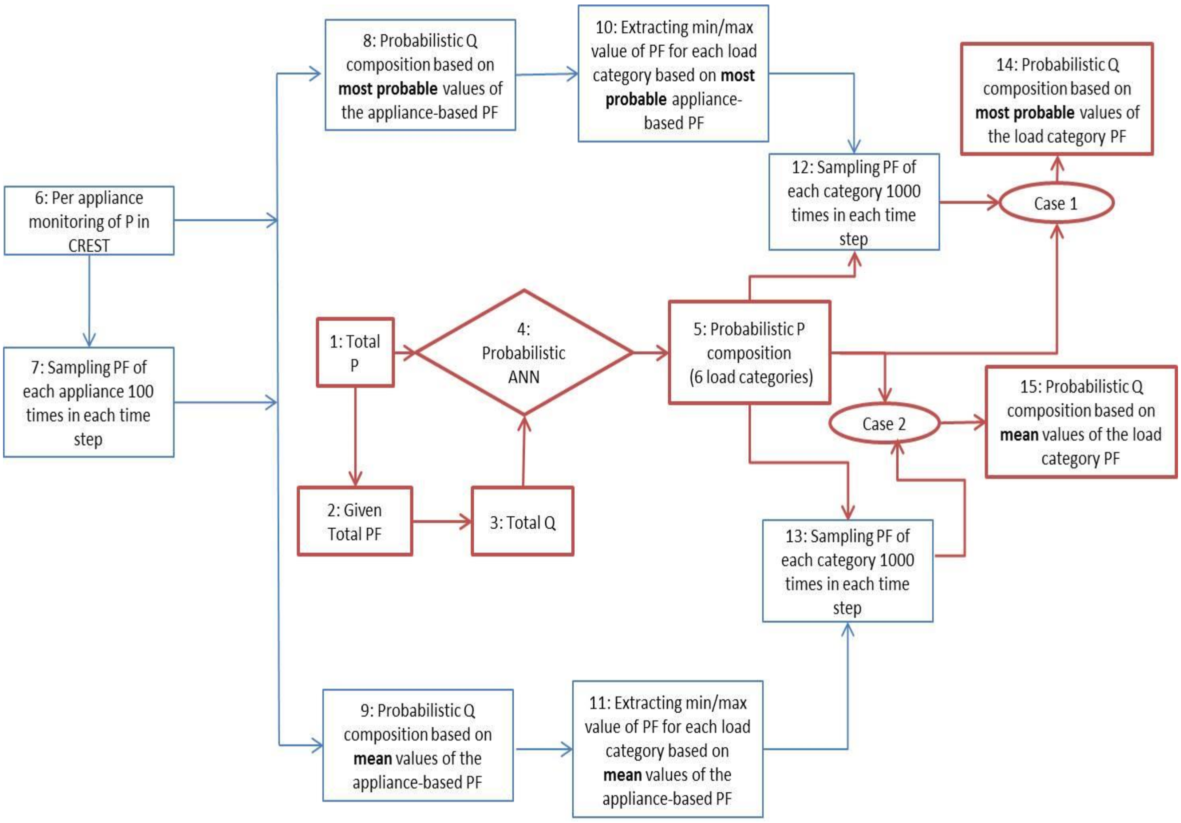

8: Probabilistic Q composition based on **most probable** values of the appliance-based PF
10: Extracting min/max value of PF for each load category based on **most probable** appliance-based PF
14: Probabilistic Q composition based on **most probable** values of the load category PF
6: Per appliance monitoring of P in CREST
12: Sampling PF of each category 1000 times in each time step
Case 1
7: Sampling PF of each appliance 100 times in each time step
1: Total P
4: Probabilistic ANN
5: Probabilistic P composition (6 load categories)
15: Probabilistic Q composition based on **mean** values of the load category PF
Case 2
2: Given Total PF
3: Total Q
13: Sampling PF of each category 1000 times in each time step
9: Probabilistic Q composition based on **mean** values of the appliance-based PF
11: Extracting min/max value of PF for each load category based on **mean** values of the appliance-based PF

Figure 3.10 Flowchart for the validation of probabilistic reactive load curve

In the next step, after active demand composition was obtained with respect to load categories (*box 5*), a range of min/max PF values for each category was derived to be used in subsequent Monte Carlo simulations, as there was no per-appliance data for this pilot site. The range was established using the CREST model [111] for 1000 end-users (*box 6*), by extracting two probabilistic values (from a range of 100 randomly generated values – *box 7*) for each appliance in each time step (relying on (3.3)), namely, the mean and the most probable value of reactive load (*boxes 8* and *9*). It should be noted that, in the presence of sub-metering data (per-appliance active demand measurements), actual measurements would be used instead of the CREST tool. Finally, by aggregating reactive load of corresponding appliances into categories, two sets of PFs per load category were derived, one based on the mean and the other based on the most probable value of reactive load. This correspondingly resulted in two ranges of PFs (min/max value) for each load category (*boxes 10* and *11*).

In the following step, PFs were randomized (following uniform distribution) 1000 times for each load category and in each time step over the two obtained ranges (*boxes 12* and *13*). Accordingly, two decomposed reactive load curves (*boxes 14* and *15*) were derived based on the active load measurements of the test site and the probabilistic PF values, using ether the most probable value (*box 14*) or the mean value (*box 15*).

The obtained reactive load curves, representing the sum of reactive load of 6 individual load categories, are presented in Figure 3.11 and compared with the original reactive load curve. The grey area shows the range between maximum and minimum possible reactive load, based on minimum and maximum PF values, respectively. It can be seen that the reactive curve built up based on mean values of the PF is closer to the actual one. The MAPE, defined in (3.4) was used to assess the accuracy of estimation, giving a 12.9% error for the curve based on the most probable PFs, and 5.3% for the curve based on mean values of the PF.

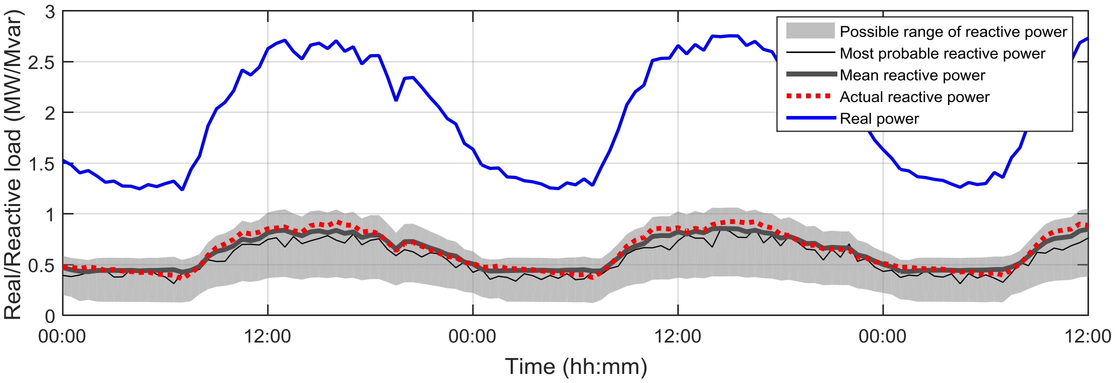


Figure 3.11 Derived and actual reactive load curves

Another measure of accuracy is the mean square error (MSE [195]) whose value was 0.0099 Mvar^2 for the curve based on the most probable PFs, and 0.0017 Mvar^2 for the curve based on mean values of the PF. The error was calculated as in (3.5).

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{Q_i - Q_{ai}}{Q_{ai}} \right| \qquad (3.4)$$

$$MSE = \frac{1}{n}\sum(Q_i - Q_{ai})^2 \tag{3.5}$$

where Q_i is the estimated value of the reactive load at a time step i, Q_{ai} is the actual value of the reactive load at the time step, and n is the number of samples. Therefore, in the equation (3.3) of the methodology, the mean value of the PF calculated from the set of probable values was used to derive the reactive load data and corresponding reactive power daily loading curve.

3.5 Artificial neural network based demand decomposition with limited demand observability

Following the flowchart shown in Figure 3.1, aggregated SM data from the monitored end-users are used for training the two-layer feed-forward ANN. The ANN is trained using total measured active and reactive power as input data and calculated participation (shares) of the six categories as the target data (block {3} in Figure 3.1). In addition to missing data restoration, and as a part of data pre-processing, the data scaling was performed in order to set all the input values in a comparable range; therefore, active and reactive load data was scaled to the range {0,1} taking maximum monthly active load of the aggregation as the base value.

The training process is performed using 7 days data (denoted as $PTRN$), which includes minute-based real and reactive power measurements, giving $7 * 1440 = 10080$ samples in total for each of the variables. The training data are presented in a matrix form as follows:

$$PTRN = \begin{bmatrix} \mathrm{P}_1 & \cdots & \mathrm{P}_\mathrm{i} & \cdots & \mathrm{P}_{7*1440} \\ \mathrm{Q}_1 & \cdots & \mathrm{Q}_\mathrm{i} & \cdots & \mathrm{Q}_{7*1440} \end{bmatrix} \tag{3.6}$$

The target data represent the participation of each load category in the total demand. If in a time step i, active load of category j equals P_{ji}, then the participation or weighted factor (WF) w_{ji}^P (in per unit) of that category is given as:

$$w_{ji}^P = \frac{P_{ji}}{P_i} \tag{3.7}$$

where P_i is the total active demand in a time step i. It is worth mentioning that in each time step the following condition has to be fulfilled (N is the total number of load categories, here equal to 6):

$$\sum_{j=1}^{N} w_{ji}^{P} = 1 \tag{3.8}$$

Target data (denoted as $TTRN$) can then be presented in a matrix form as follows:

$$TTRN = \begin{bmatrix} w_{1,1}^{P} & \cdots & w_{1,7*1440}^{P} \\ w_{2,1}^{P} & \cdots & w_{2,7*1440}^{P} \\ \vdots & \dots & \vdots \\ w_{6,1}^{P} & \cdots & w_{6,7*1440}^{P} \end{bmatrix} \tag{3.9}$$

In case of the reactive power, the participation of each category is calculated as follows:

$$w_{ji}^{Q} = \frac{Q_{ji}}{Q_i} = \frac{P_{ji}\tan(\varphi_{ji})}{P_i\tan(\varphi_i)} = w_{ji}^{P}\frac{\tan(\varphi_{ji})}{\tan(\varphi_i)} = w_{ji}^{P}\frac{\left(\frac{\sqrt{1-PF_{ji}{}^{2}}}{PF_{ji}}\right)}{\left(\frac{\sqrt{1-PF_{i}{}^{2}}}{PF_{i}}\right)} \tag{3.10}$$

where φ_{ji} and φ_i are phase angles of category j and total load in time step i, respectively, and PF_{ji} and PF_i are corresponding power factors.

A two-layer feed-forward ANN with Bayesian Regulation Backpropagation, similar to the one introduced in [196], was chosen for load decomposition due to its robustness and satisfactory accuracy reported in [39]. The data in [39] was generated probabilistically, without any measurements available. Therefore, in order to assess, under the same conditions, the improvement in accuracy of demand decomposition with the inclusion of per-appliance measurement data, the ANN settings were not changed. This ANN is a two layer neural network, with one input, one hidden and one output layer, where the input layer has two neurons (for total active and reactive load inputs), and the output has six neurons – one neuron representing the share of a category. The transfer functions of the hidden and output layer are log-sigmoid and tan-sigmoid, respectively, as suggested in [196]. The sigmoid functions are chosen for the transfer functions as the expected output of the ANN (the shares of different load categories) is in the range of $[0,1]$. Even though the hidden layer transfer function is log-sigmoid, which limits the output to the range of $[0,1]$, the output layer transfer function is tan-sigmoid and its output can range from -1 to 1, hence it provides higher sensitivity to its input values. As the number of the training samples ($N = 7 * 1440$) is much larger than the number of input variables ($d = 2$), the number of neurons in the hidden layer (n) is calculated as follows [16]:

$$n = \sqrt{\frac{N}{d\ lnN}} \tag{3.11}$$

Once trained, the ANN (block {4} in Figure 3.1) uses total active and reactive demand forecast at the aggregation (bulk) point (block {5}) as the input, giving its load composition as the output (block {6}). Finally, forecast demand composition of both monitored and non-monitored end-users is obtained. It should be noted that for real-time applications of demand decomposition, real time (measured) values of total active and reactive demand at the bulk point would be used as input to the trained ANN instead of forecast values.

Figure 3.12 illustrates the input and output for the ANN, while Figure 3.13 represents the architecture of the ANN used in Matlab [128], where *w* and *b* correspond to weights and biases of the network, respectively, assigned to the inputs in hidden and output layer. The number of hidden layer neurons is 23 in this case, as the values for N and d in (3.11) are $7 * 1440$ and 2, respectively.

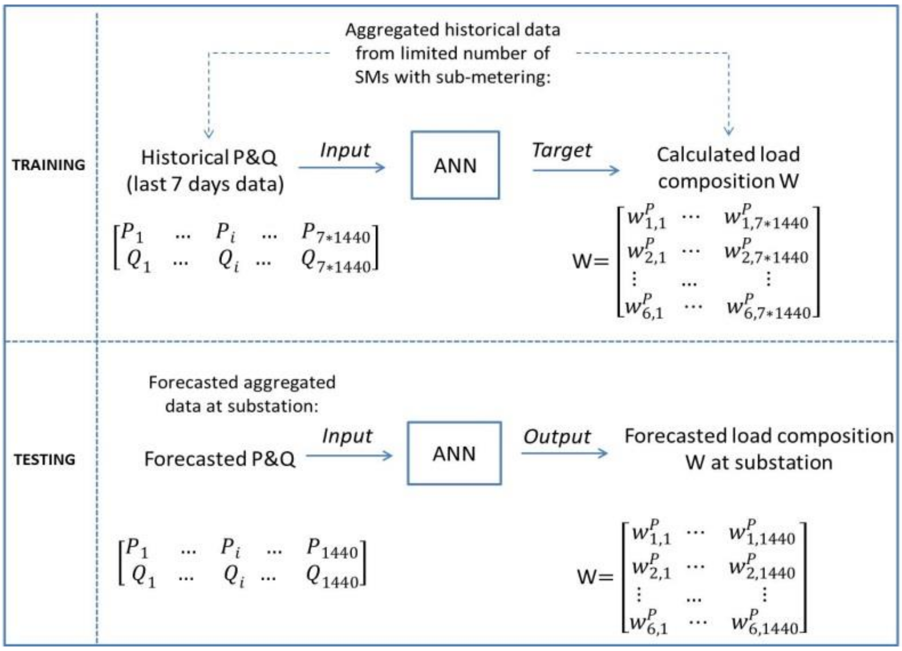

TRAINING
Aggregated historical data from limited number of SMs with sub-metering:
Historical P&Q (last 7 days data)
Input
ANN
Target
Calculated load composition W
$\begin{bmatrix} P_1 & \dots & P_i & \dots & P_{7*1440} \\ Q_1 & \dots & Q_i & \dots & Q_{7*1440} \end{bmatrix}$
$W=\begin{bmatrix} w^P_{1,1} & \cdots & w^P_{1,7*1440} \\ w^P_{2,1} & \cdots & w^P_{2,7*1440} \\ \vdots & \dots & \vdots \\ w^P_{6,1} & \cdots & w^P_{6,7*1440} \end{bmatrix}$
TESTING
Forecasted aggregated data at substation:
Forecasted P&Q
Input
ANN
Output
Forecasted load composition W at substation
$\begin{bmatrix} P_1 & \dots & P_i & \dots & P_{1440} \\ Q_1 & \dots & Q_i & \dots & Q_{1440} \end{bmatrix}$
$W=\begin{bmatrix} w^P_{1,1} & \cdots & w^P_{1,1440} \\ w^P_{2,1} & \cdots & w^P_{2,1440} \\ \vdots & \dots & \vdots \\ w^P_{6,1} & \cdots & w^P_{6,1440} \end{bmatrix}$

Figure 3.12 Detailed presentation of ANN input and output, during training and testing process

In order to improve the accuracy of the ANN output, the network training can be repeated regularly, as the measurement (input) data gets updated. For example, the input can be updated every 6 hours with the most recent 7 days data and any historical data older than that can be discarded. The results of ANN-based load disaggregation and the influence of missing data and the level of SM coverage will be presented in the following sections.

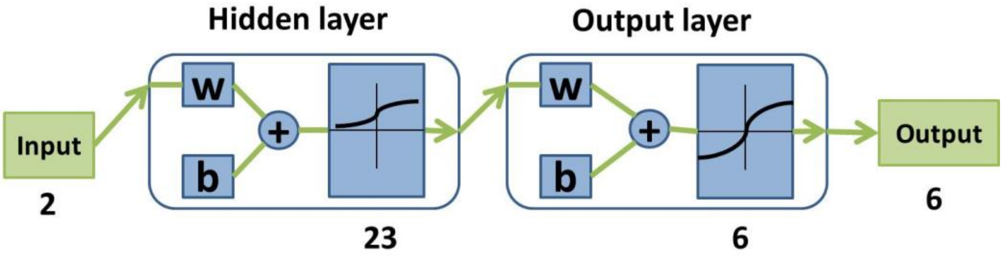


Figure 3.13 Architecture of the feed-forward network used in Matlab, reproduced from [128]

3.5.1 Case studies

In order to test the accuracy of the ANN, the network is tested using one day (the eighth day) data from the historical dataset, which served as the day-ahead load forecast. The effect of larger training data, e.g., 28 days, as well as the effect of the inaccuracy of the total load forecast on the accuracy of demand decomposition is discussed in Section 3.6. The training of the ANN with 7 days data took between 3 and 8 minutes using a PC with the 64-bit operating system and 3.40 GHz processor. Once the ANN is trained, the forecast active and reactive load curves at the substation (block {5} in Figure 3.1) are used as the input to the trained ANN (block {4} in Figure 3.1). The output is presented in the form of the decomposed forecast active and reactive load curves for all aggregated end-users (monitored and non-monitored ones), similarly to the representation in Figure 3.2. It is important to note that the training dataset should not include days with activated DR programs, as in this case the data would not show the actual DR potential (before the DR action), but the loading curve and demand composition after load shifting or curtailment.

An aggregation of 1000 households is analysed, illustrating a relatively high number of users. The CREST load model [111] was used to generate individual load profiles (decomposed into home appliances) over one month, which also served as training and testing data sets for the ANN. With CREST model it is possible to generate numerous, statistically proven, daily load curves (for each appliance in a household) based on the month of the year (in this example January was chosen), number of residents per household, and type of the day (working/non-working). The residential occupancy statistics for the UK (29% of households accommodate a single resident, 35%

accommodate two, 16% have three residents and 20% have four) is adopted from [197] to generate appropriate load profiles of residential customers.

The three illustrative case studies analysed in this section are as follows:

- *Case A: Smart metering system with all meters sending data every minute, with no missing data;*
- *Case B: Smart metering system with meters sending data every minute, with 20% missing data ("NaN" values);*
- *Case C: Smart metering system with different meters sending data at different time steps (1, 10, 30 and 60 minutes) and with 20% missing data.*

In order to assess the required percentage of users with sub-metering data for confident demand decomposition, five levels of SM coverage are investigated within each case study: 5% coverage (50 households out of 1000 have SMs with sub-metering technology), 10%, 20%, 50% and 80% SM coverage. The ANN is trained, correspondingly, with sub-metering data coming from 50, 100, 200, 500 or 800 households. The objective of these examples is to illustrate the effect of SM (with sub-metering functionality) coverage on the accuracy of demand decomposition in an aggregation of 1000 households, as well as the effect of missing data. The accuracy is assessed based on the composition of the aggregated load (during the eighth day) obtained from the actual values in the given dataset with 1000 households. Absolute weighted factor error (AWFE) is used for this purpose, and calculated at each time step (minute) as follows:

$$AWFE_{cat} = \left| WF_{cat,\,ANN} - WF_{cat,real} \right| \tag{3.12}$$

where $WF_{cat,\,ANN}$ is the share of the load category obtained as the result of the ANN, and $WF_{cat,\,real}$ is the actual share of the category, both given in p.u. based on the average aggregated monthly active demand.

Figure 3.14 illustrates the way errors are accounted for, on the example of two load categories, namely controllable and uncontrollable load. If the total load at time t equals 0.7 p.u. (where 1 p.u. refers to the average monthly load at the aggregation point, which is in this study around 0.6 MW), and the estimated load shares of controllable and uncontrollable load are 0.3 p.u. and 0.4 p.u., respectively, then the real values of the shares are within the following ranges: $P_C = 0.3\, p.u. \pm AWFE$ and $P_{UC} = 0.4\, p.u. \pm AWFE$.

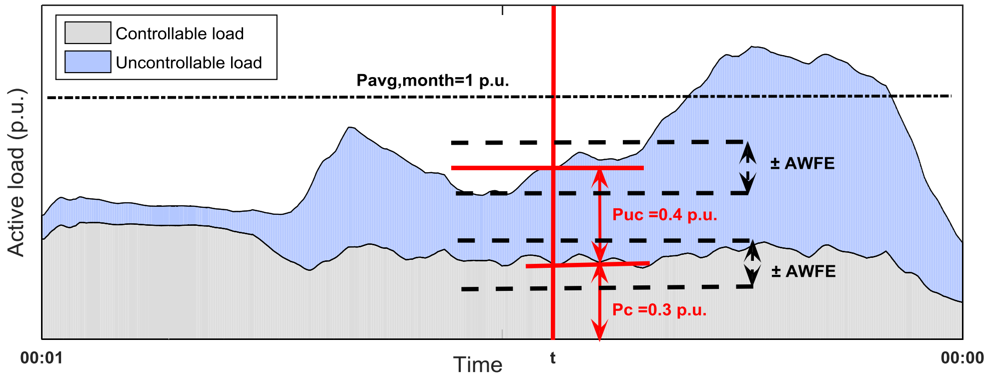


Figure 3.14 Presentation of the confidence level on the example of controllable/uncontrollable load over one day (24 hours)

In all the cases (A to C), the errors are compared to those obtained with 100% SM coverage, as the reference case. In this way the error coming from the ANN itself is revealed. As previously mentioned, the load forecasting error is not taken into account in this analysis – it is addressed in Section 3.6.2. In addition, the errors are compared with those in case of 0% SM coverage (no SMs installed at the users' premises), where the ANN is trained with probabilistically derived data, originating from statistical data about the electricity usage in UK domestic sector [194]. According to these data, controllable load within the total daily load ranged between 15% and 50%. The training data was generated following approach described in [39]. At the 0% coverage, the same ANN, trained with probabilistically generated data, is used in all three considered cases (A, B and C). Therefore, the accuracy of demand decomposition is the same in the cases with 0% SM coverage level.

The results of the analysis are presented in the form of cumulative density functions (CDFs) of the AWFE over a range of SM coverage levels, including the reference (100%) and 0% coverage, for controllable load shares only. At the same time, the 90th percentile confidence level of the AWFE for different load categories and controllable load are presented in a form of bar plots over a range of SM coverage levels. The 90th percentile is chosen as it shows the maximum error for 90% of the observed time steps (here, 1296 out of 1440 time steps over a 24h period).

3.5.1.1 Active load decomposition

The accuracy of estimation of controllable load shares is very similar between cases A-C, which is why only case C is presented in Figure 3.15, as the one with the highest share of missing data before pre-processing. It can be seen from the figure that all SM coverage levels provide errors smaller than 0.1 p.u. (i.e., 10% of the average monthly load, which corresponds to around 60 kW) in 90% of the time steps. In cases where there is no sub-metering provided and the estimation can only be done probabilistically (0% SM coverage), the 90th percentile of the AWFE is 0.23 p.u, which corresponds to around 140 kW. It can be also seen that for the SM coverage levels of 50% and higher, the accuracy remains the same. The calculated errors for the three cases (A to C) do not change notably, confirming that 20% missing data, and different sampling steps, do not affect the accuracy significantly if the missing samples are restored. This also confirms that the use of simple data restoration method (here, LI) is fully justified.

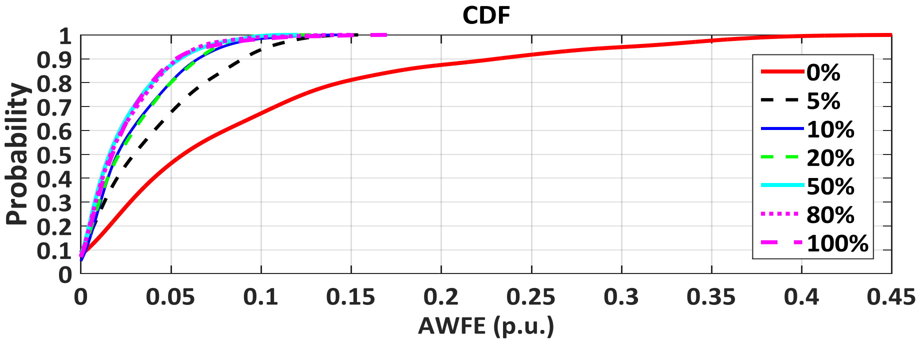


Figure 3.15 CDF of AWFE for the estimation of controllable active load, case C

Figure 3.16 illustrates the load composition in the base case and with 5% SM coverage. It should be noted that the total active demand is the same in the two figures, only the shares of the load categories differ. Finally, Figure 3.17 presents the shares of individual load categories in the two cases, i.e., based on actual values and based on ANN trained with data from 50 users. The figure shows that most categories are well estimated, except categories CTIM1 and Lighting, which show the highest discrepancies. It should be noted that CTIM1 and QTIM1, as controllable load categories, have lower shares in the daily load curve compared to R_C, as another controllable load category. The main reason for this is the fact that the observed dataset represents demand in January, when heaters are used more than loads modelled as induction motors.

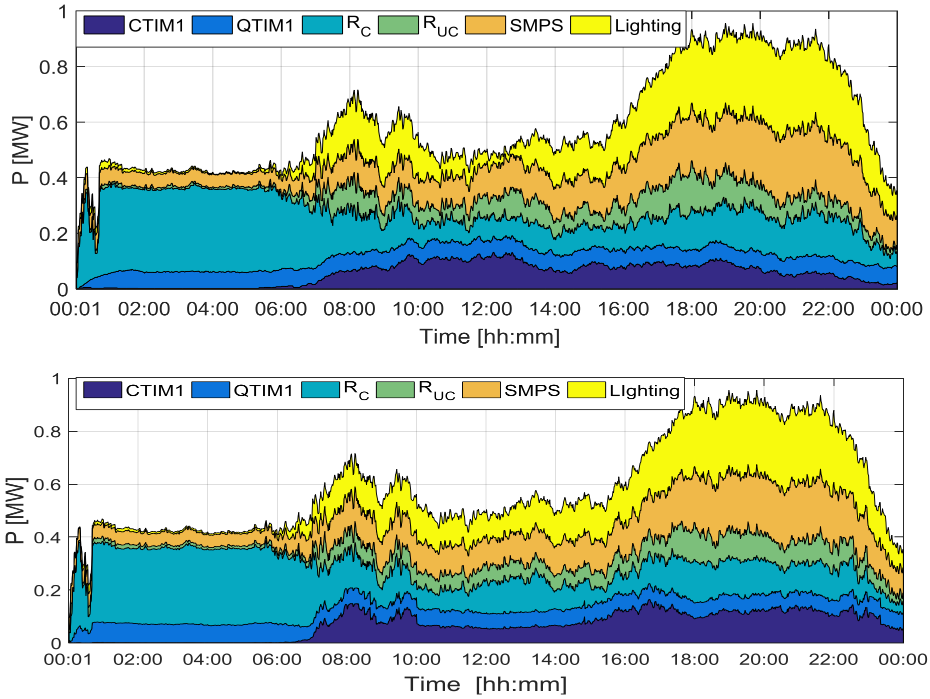


Figure 3.16 Demand composition of 1000 end-users' active load based on actual values (top) and based on the ANN trained with data from 50 users (bottom)

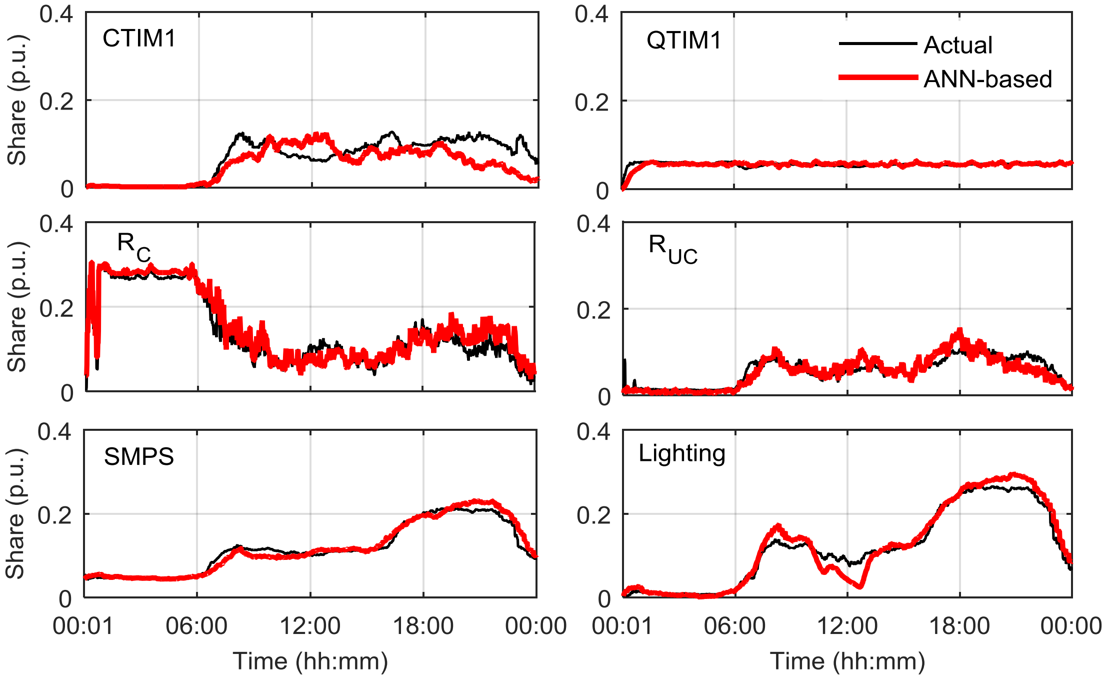


Figure 3.17 Shares of individual load categories based on actual values and those estimated using ANN trained with data from 50 users

The estimated shares of individual load categories show similar accuracy (errors up to 0.1 p.u.) in most cases, as seen in Figure 3.18, which represents the 90th percentile of AWFE read from the CDF plots for each load category in case C. The highest accuracy is seen with QTIM1, and the lowest with CTIM1. The shares of some categories, namely CTIM1 and QTIM1, are very accurately assessed even with 0% SM coverage (with errors around 0.1 and 0.05 p.u., respectively), which shows that in these cases only statistical data is sufficient for confident load decomposition. For all other categories, probabilistic approach introduces higher errors. Different accuracy in prediction of participation of different load categories in total demand can be attributed to variation of particular load category during the observed period.

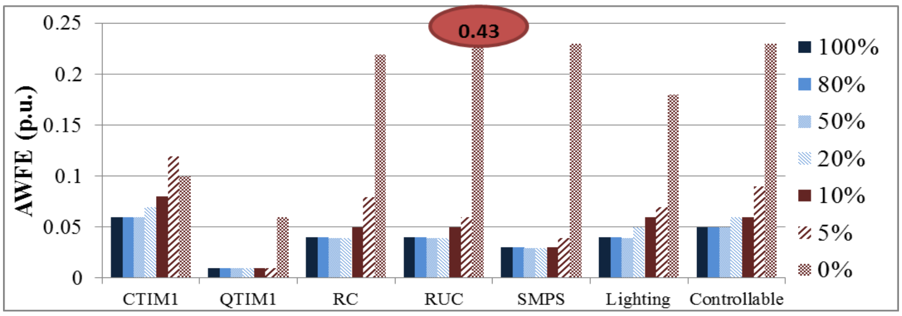


Figure 3.18 AWFE with 90th percentile confidence level for real power, case C

The correlation between total active/reactive demand and shares of different load categories is studied using Spearman's rank coefficient of nonlinear correlation [198], as the correlation between the parameters is not linear. The coefficient is shown in Table 3.4 and **Error! Reference source not found.** over the period of one week and one month, respectively, for the case C and three aggregation levels.

Table 3.4 Spearman's coefficients between the total active/reactive load and shares of different load categories for different aggregation levels over the training period (one week), case C

Total load	Aggregation level	CTIM1	QTIM1	Rc	Ruc	SMPS	Lighting
P	1000 users	0.42	-0.89	-0.50	0.53	0.49	0.79
	200 users	0.36	-0.89	-0.34	0.44	0.15	0.67
	50users	0.34	-0.87	0.36	0.25	-0.46	0.47
Q	1000 users	0.49	-0.82	-0.66	0.56	0.67	0.92
	200 users	0.51	-0.77	-0.60	0.51	0.44	0.87
	50users	0.54	-0.71	-0.02	0.34	-0.11	0.76

Table 3.5 Spearman's coefficients between the total active/reactive load and shares of different load categories for different aggregation levels over one month, case C

Total load	Aggregation level	CTIM1	QTIM1	Rc	Ruc	SMPS	Lighting
P	1000 users	0.40	-0.89	-0.51	0.51	0.50	0.78
	200 users	0.31	-0.88	-0.30	0.39	0.26	0.65
	50users	0.25	-0.84	0.26	0.22	-0.30	0.44
Q	1000 users	0.48	-0.81	-0.68	0.55	0.70	0.92
	200 users	0.49	-0.74	-0.60	0.50	0.57	0.87
	50users	0.49	-0.65	-0.18	0.34	0.11	0.78

It can be seen that the correlation is similar for both periods, confirming that there is no need for larger historical data to be used for ANN training. Category QTIM1 shows the highest Spearman's coefficient (Spearman's rho) at most aggregation levels, followed by lighting, hence the high accuracy in prediction. Category CTIM1, on the other hand, has lower Spearman’s coefficient than other categories in general, hence lower correlation with total active and reactive load, which is leading to the lowest accuracy in prediction of the share of CTIM1 in the total load.

3.5.1.2 Reactive load decomposition

The estimation of the shares of the controllable reactive load is also very accurate, with the 90th percentile AWFE between 0.04 p.u. and 0.08 p.u. over the range of SM coverage levels, as shown in Figure 3.19. In case of 0% SM coverage, the 90th percentile of AWFE is 0.17 p.u., which corresponds to around 100 kvar. Similarly to active load, there is only a minor deterioration in the accuracy in cases B and C at lower SM coverage levels, compared to case A.

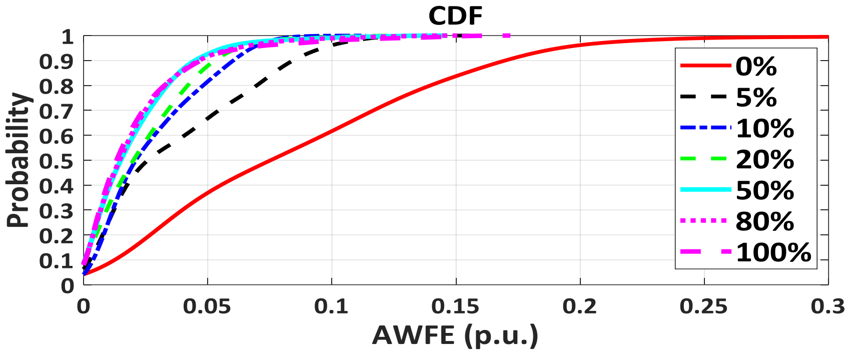


Figure 3.19 CDF of AWFE for the estimation of controllable reactive load, case C

The estimation of the shares of the load categories results in 90th percentile of AWFE between 0 and 0.08 p.u. for all the SM coverage levels, as shown in Figure 3.20. The reactive power of the controllable resistive loads (R_C) equals zero in all time steps, which is why the 90th percentile of AWFE is around 0, except for the case of probabilistic approach. The same applies to uncontrollable resistive loads (R_{UC}) which are, due to some home appliances, such as oven, modelled as imperfect resistors with power factor lower than 1. Except for these two categories, the QTIM1 share is estimated with the highest accuracy and the CTIM1 and Lighting with the lowest. Figure 3.21 illustrates the reactive load composition in the base case and in the case with 5% SM coverage.

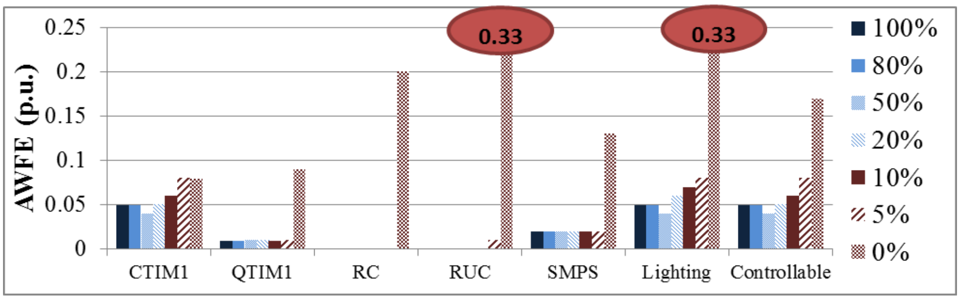


Figure 3.20 AWFE with 90th percentile confidence level for reactive power, case C

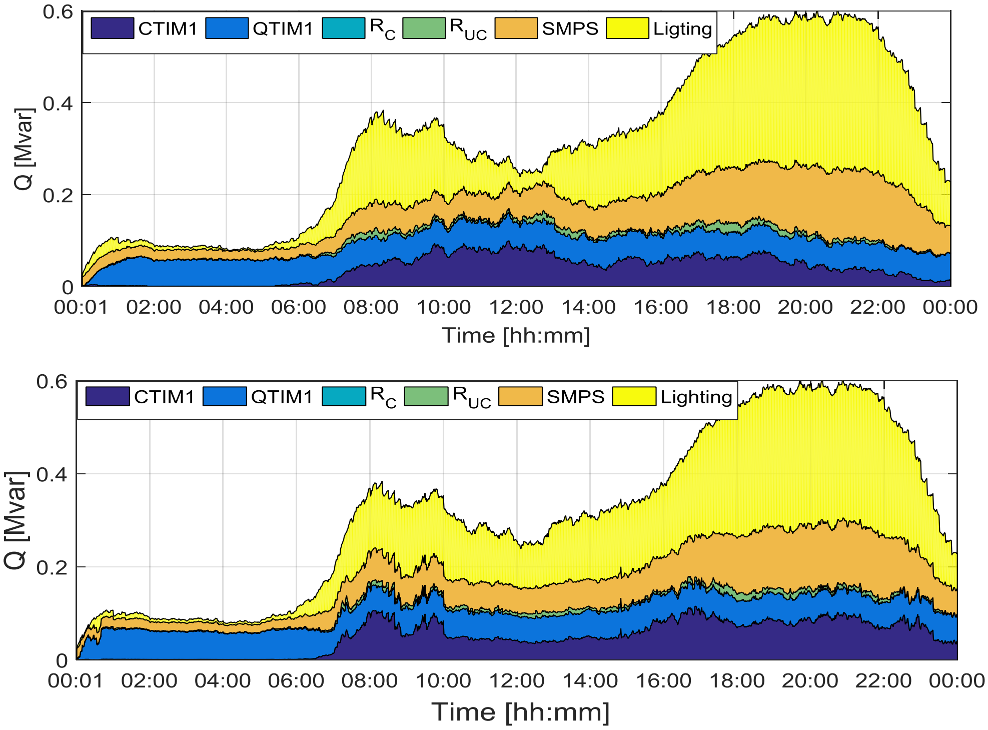


Figure 3.21 Demand composition of 1000 end-users' reactive load based on actual values (top) and based on the ANN trained with data from 50 users (bottom)

3.5.1.3 Discussion

The results of the case studies have shown the effect of missing data and different SM coverage on the accuracy of forecast active and reactive load composition. Based on the studies performed and illustrative results shown in the previous section, it can be calculated that the overall accuracy of the assessment is not significantly affected by SM coverage, nor by missing data. The accuracy, though, changes more with SM coverage level, than with missing data. Furthermore, the daily load shares of some load categories (CTIM1 and QTIM1), can be estimated very accurately using statistical data of electricity usage in the area only, i.e., without any SMs. For other load categories, the utilization of ANN with probabilistically generated training data is justified only if the target application of demand decomposition does not require high accuracy. It was also observed that the estimation of reactive load composition with 0% SM coverage is more accurate than the estimation of active load.

Furthermore, for the DR programs relying on wet appliances (CTIM1) and cold appliances (QTIM1), limited statistical data are sufficient to estimate their load shares. At lower SM coverage levels, AWFE values for these two categories are up to around 0.1 p.u. In case of voltage-based DR programs (namely conservation voltage reduction) relying on resistive loads (e.g., water heaters), AWFE of the load share equals 0.06 p.u. at 5% SM coverage, while it is lower than 0.05 p.u. for other SM coverage levels.

For most of the load categories even the lowest SM coverage (5%) enables very high accuracy of identification of load composition. The results show that even with minimal investments in sub-metering technologies (for only 5% of the users) the desired accuracy of load composition forecast can be obtained, and notably improved compared to the probabilistic approach, when no sub-metering data is available (0% SM coverage). The minimum SM coverage and resulting accuracy of estimation, however, may be different for different applications, and needs to be investigated further.

Even though the methodology has been illustrated on residential load sector, it can be equally well applied to other load sectors, e.g., commercial, industrial or mixed. Finally, it is important to note that the proposed methodology is area-dependent. Therefore, the accuracy is higher when the data used for training the ANN comes from the electrical or geographical “neighbourhood” of the aggregated users. Finally, if only statistical data is available for ANN training, with no sub-metering data coming from individual end-users, it has to correspond to the type of the users under analysis, as the daily range of controllable load differs among industrial, commercial and domestic types of users.

3.5.2 Validation of the methodology

In order to validate the methodology on real data, another dataset was chosen, namely the Pecan street electricity consumption data [24]. These data include minute-based measurement data from 200 residential users located in Austin, Texas, during the three month period between June-August 2015. All of the observed time series had all samples available, i.e., there was no missing data, similarly to case A described in Section 3.5.1. Therefore, only the influence of SM coverage was examined, as a more influential factor on the accuracy of demand decomposition, as discussed in the previous section. The sub-metering data are also available in the database covering a number of years, starting from 2013. These include circuit-based measurements, e.g., total consumption of the living room, bathroom, or kitchen plugs, and in some cases measurements of individual appliances (washing machine, furnace, etc.). Therefore, some assumptions on the load categories used in different parts of the household were made to enable load classification into 6 categories, similarly to Table 3.1. For example, it was assumed that half of the loading in the living room consisted of lights and half consisted of the electronic appliances (SMPS category). Weather data included temperature, humidity and wind speed measurements from the corresponding period of the year, together with the day type (1 for working day and zero for non-working day). The dataset did not contain reactive load data, so the probabilistic approach was taken to derive this, following steps described in Section 3.4. The ANN trained with probabilistic data only (with 0% SM coverage) was not taken into account in this study.

Figure 3.22 presents the CDF of AWFE for controllable active load estimated for one day in August based on ANN trained with 7 days historical data. The range of the 90th percentile of AWFE is very similar to the ones in the test cases presented in Section 3.5.1, which confirms the validity of the methodology and the fact that the training data corresponding to the observed set of consumers yields high accuracy.

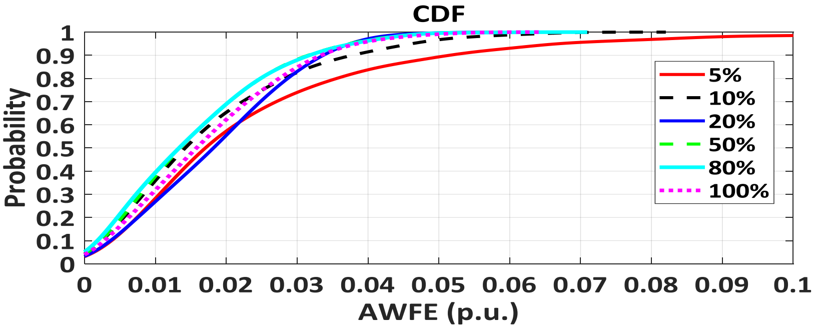


Figure 3.22 CDF of AWFE for the estimation of controllable reactive load for one day in August

Figure 3.23 and Figure 3.24 illustrate the forecast active load composition in case of 5% SM coverage and in the base case (actual measurements), respectively. Clearly, category CTIM1 has the highest share, as it mainly consists of HVAC units, which are highly used in Texas. The second highest share belongs to controllable resistive loads (R_C), namely water heaters. As both of these categories are controllable and consist of thermostatically controlled loads, it can be concluded that there is a large DR potential from the residential users in this area.

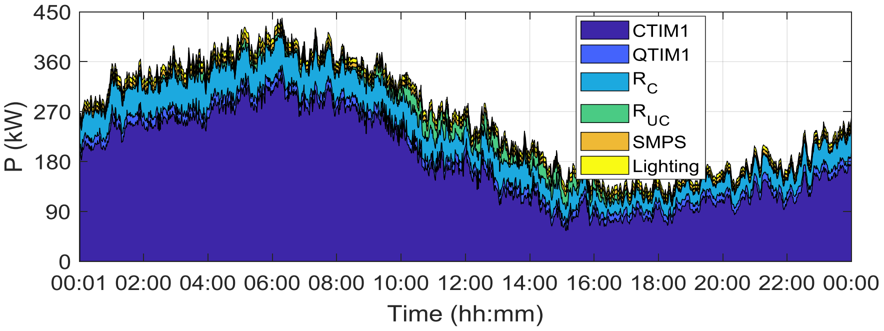


Figure 3.23 Estimated active demand composition of 200 customers with 5% SM coverage (10 customers monitored)

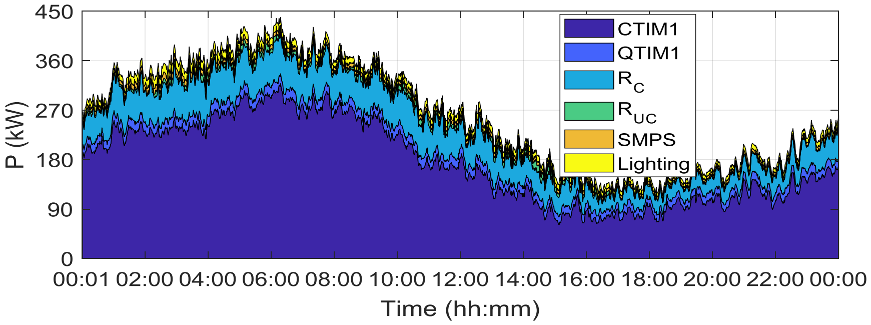


Figure 3.24 Active demand composition of 200 customers based on field measurements

In order to verify the methodology on a longer period of the year, and at the same time reduce the computational burden, the approach was tested on three consecutive months of the Pecan street dataset, namely June, July and August. This period was chosen to illustrate end-users' behaviour during the summer, when there is usually a significant change in the load between consecutive months due to the summer holidays. The simulations were done iteratively, by estimating the load composition of 200 users every day, based on the last 7 days training data. This means that the ANN was retrained every day with the most recent 7 days' data. Therefore, the size of the training data is always the same, but updated every day with the newest 1440 samples (the "oldest" 1440 samples are discarded in each update). This process was done for each SM coverage level individually, and tested versus the base case (actual measured data). It should be noted that even with 100% SM coverage, there is only information about the demand composition of the previous days, while for the next day only total active and reactive load forecast is given as the input to the trained ANN.

Figure 3.25 presents the AWFE for the forecast of the share of controllable load for different SM coverage levels. The range of errors equals the one presented for one day testing (illustrated in Figure 3.15). The effective mismatch in the size of controllable load (in kW) is slightly smaller here, due to the smaller base value. The per unit values of the errors are obtained based on the mean load during the 3 observed months, which was 212.9 kW, giving a maximum mismatch of 12.8 kW (AWFE = 0.06 p.u.) in 90% of the cases, at 5% and 10% SM coverage. The maximum possible mismatch in most cases at 20%, 50% and 80% SM coverage equals 8.5 kW (0.04 p.u.). In case of 100% SM coverage, the 90th percentile AWFE is 0.03 p.u. (6.4 kW mismatch). If a realistic assumption is made, based on the observed customers, that the average consumption of a residential air-conditioning (AC) unit is around 1kW, this means that the maximum mismatch of the size of controllable load at the aggregation level of 200 users is around 13 AC units for the lowest SM coverage levels, and around 6 AC units for the highest SM coverage level.

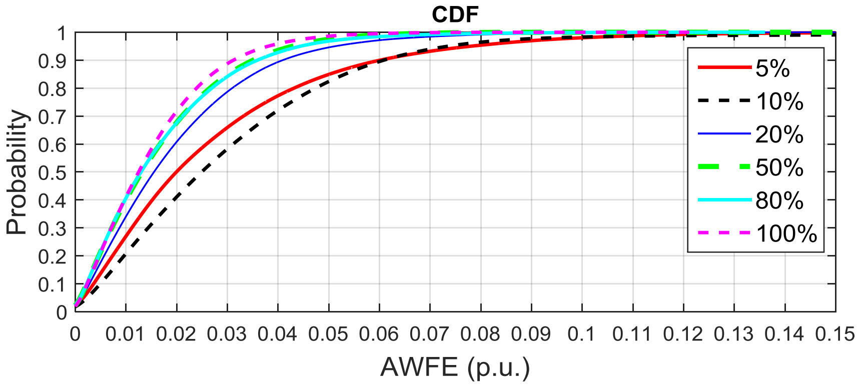


Figure 3.25 CDF plot of AWFE for controllable load when testing the approach on three consecutive months of the Pecan street dataset

3.5.2.1 Discussion

As already mentioned, the results are area-dependent, i.e., they depend on the users' daily habits in the observed area. To ensure high accuracy in prediction even for the same aggregation of users, it is advised to repeat the training of the ANN by updating the training data with the most recent historical data. The reason for this lies in high variability in daily consumption, of individual users in particular [199], that happens due to the difference in weather, season, holidays, etc. Similarly, for other types of load sectors, such as industrial, commercial, or mixed, corresponding measurement/statistical data should be used for training the ANN and updated with a certain resolution, e.g., every week. Methodology used to obtain the results, however, would not have to be changed at all, if different types of customers, different geographical locations or time periods or seasons are considered. The only thing that needs to be accordingly updated is the training data set.

3.6 The effect of input data on the accuracy of demand decomposition

This section analyses how different factors may influence the accuracy of demand decomposition, namely the size of training data and the accuracy of total demand forecast at the aggregation (substation) level.

3.6.1 The effect of weather data and the size of training data

Weather and type of the day (working/non-working) are considered as factors influencing the daily loading curve, especially with respect to cooling and heating devices [200]. In order to assess the necessity of this data in the training of the ANN for demand decomposition, two cases are examined: *i*) the ANN training input consists only of total active and reactive demand, as suggested in the methodology; *ii*) the ANN training input consists of total active and reactive demand, weather data (temperature, relative humidity and wind speed) and day type values (1 for working days and 0 for weekends/public holidays). Consequently, the ANN in the latter case has 6 input parameters. The ANN training target (TTRN, as defined in (3.9)) is the same in both cases. When dealing with multifaceted data, an automatic feature selection algorithm could be incorporated to extract only the relevant data attributes for the training purposes. Based on the past experience and the number of data features considered in this case study, that was not deemed necessary, and a simple trial and error approach was used to decide on the necessity of inclusion of weather data. It should be noted that Pecan street data [24] was used in the analysis in this section.

Figure 3.26-Figure 3.28 show the CDF of the AWFE for one day in August 2015, based on 1440 samples, with the ANN trained with the data from the past two months, one month and one week, respectively. In all the figures, cases without and with weather data were compared. It can be noted that weather data does not improve significantly, if at all, the accuracy of estimation in most of the cases. Also, the use of "longer historical data" (e.g., measurements from the past two months compared to past week) does not make any improvement in accuracy, quite the contrary, the longer historic data results in slightly reduced accuracy at lower SM coverage levels. This yields the conclusion that only the most recent historical data (last 7 days) are sufficient for confident load decomposition. Finally, when the ANN is trained and tested using the pilot site data, the calculated range of the 90th percentile AWFE corresponds to the one when the CREST load model was used (as presented in Figure 3.15), which validates the proposed methodology.

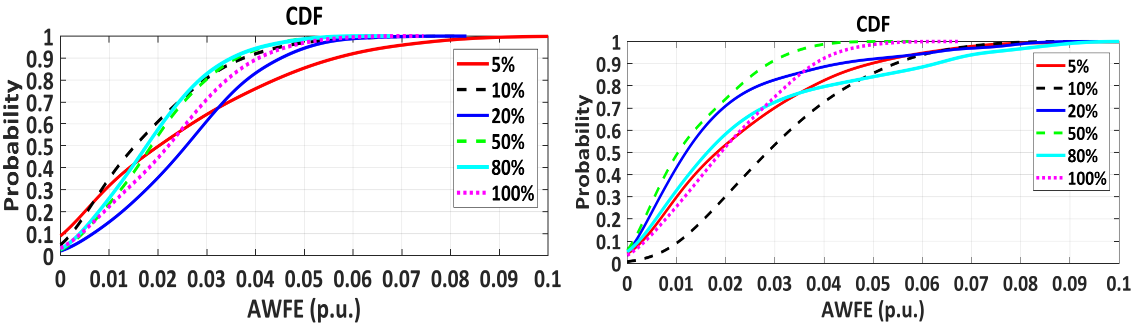


Figure 3.26 Controllable load estimation for one day in August with 60 days training (June+July) without weather data (left), and with weather data and day type (right)

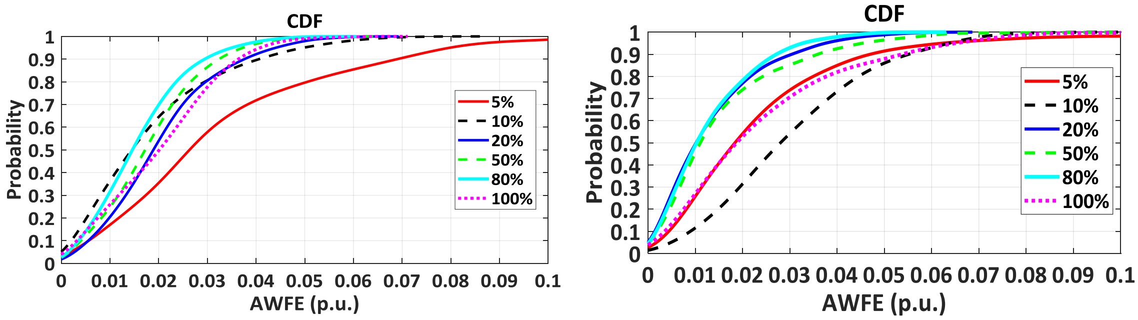


Figure 3.27 Controllable load estimation for one day in August with one month training (July) without weather data (left), and with weather data and day type (right)

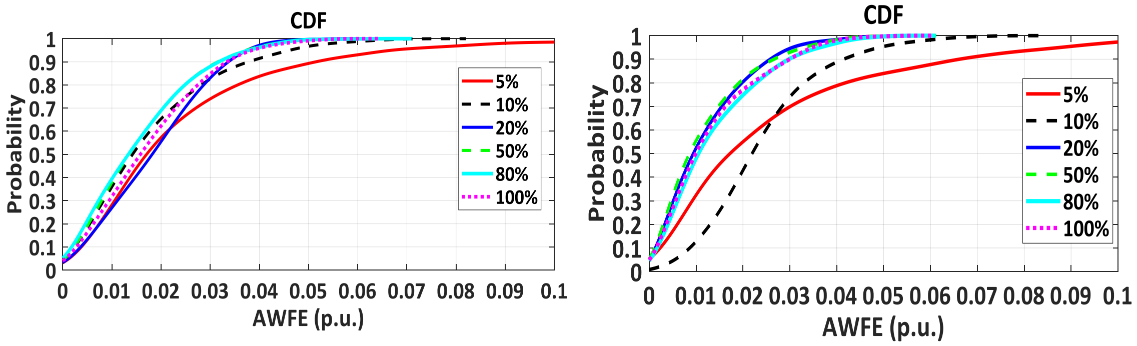


Figure 3.28 Controllable load estimation for one day in August with 7 days training (August) without weather data (left), and with weather data and day type (right)

As the mean value of the load during the observed period was 238.4 kW, if the 90th percentile AWFE of the estimated controllable load is around 0.04 p.u. (with 10-20% SM coverage in Figure 3.28 left), that means that in 90% of the cases the maximum over/under-estimation of the size of controllable load is 9.5 kW for the aggregation of 200 users. For illustration purposes, it can be assumed (based on the observed sample of consumers) that the average consumption of an AC unit is 1 kW. Thus the maximum mismatch in the size of controllable load equals to the consumption of 10 AC units at the level of 200 users for 10-20% SM coverage. 5% SM coverage usually results in slightly higher errors, around 0.06 p.u. (14.3 kW ≈ 14 AC units).

In order to clarify further the effect of the size of the training data set and the inclusion of weather data, Figure 3.29 illustrates the most probable weighting factor error (WFE):

$$\mathrm{WFE} = \mathrm{WF}_{\mathrm{cat,\,ANN}} - \mathrm{WF}_{\mathrm{cat,real}}, \tag{3.13}$$

for different SM coverage levels, size of the training data and with/without weather data. $WF_{cat,\,ANN}$ is the share of the load category obtained as the result of the ANN, and $WF_{cat,\,real}$ is the actual share of the category, both given in p.u., based on the average aggregated monthly active demand. The most probable value was extracted from the range of WFE values over the 1440 samples. The following conclusions can be deduced from the figure:

- The use of weather data (patterned bars in the diagram) improves the accuracy in cases with higher SM coverage levels, starting from 20% SM coverage (note that both positive and negative errors are taken into account), while with lower SM coverage levels (5% and 10%) the accuracy is deteriorated. This can be explained by higher randomness of aggregated load curve at lower aggregation levels.

- The use of larger training data sets mostly results in similar absolute accuracy to the one provided by smaller training data sets.

- The size of historical data (training data sets) has larger influence on the accuracy in cases of higher SM coverage levels (>20%), i.e., greater variation in accuracy can be observed with different lengths of data sets. In both cases, with and without weather data, the accuracy is slightly higher with shorter data sets.

The time required for training the network, as presented in Table 3.6, is affected by the number of neurons in the hidden layer of the ANN (calculated as $n = \sqrt{N/(d \cdot lnN)}$ [16], where N is the number of the training samples - here the number of training days times 1440, and d is the number of input variables, which is either 2 or 6 in the observed cases) – this number is lower in the case of more input variables (such as weather) for the same length of the training data (e.g., for one month of historical data). It should be noted that the simulations were run on a PC with the 64-bit operating system and 3.40 GHz processor.

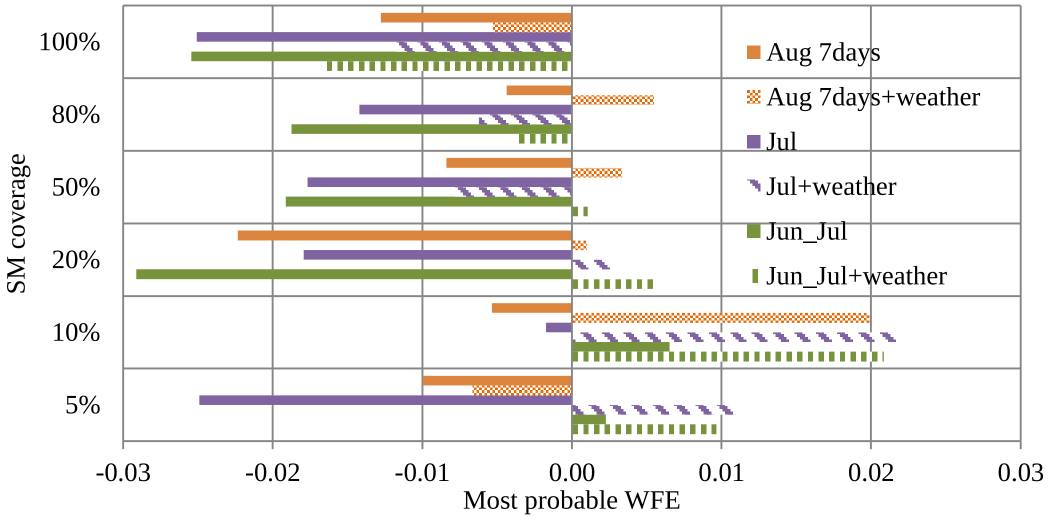


Figure 3.29 Most probable WFE based on ANN trained with different SM coverage levels, with and without weather data

Table 3.6 Training time for different size of training data

Training size	Number of hidden neurons (*n*)	Training time
2 months	62	up to 70 min
2 months + weather	36	up to 50 min
1 month	45	up to 20 min
1 month + weather	26	up to 25 min
1 week	23	up to 2 min
1 week + weather	13	up to 80 seconds

3.6.2 The effect of total demand forecasting error

This subsection illustrates the influence of the total load forecasting error in the input data used for ANN training (presented in Figure 3.12) on the demand decomposition accuracy. Total load of 200 users from the Pecan street dataset [24] was thus forecast for August 30th 2015 using an ANN trained with historical loading and weather data during the period from June 1st to August 29th 2015. This ANN, used solely for total active/reactive demand forecast at the aggregation level, had the same settings as the one used for load decomposition, but with different input and target, as shown in (3.14) and (3.15), where $P_1 \ldots P_n$ are the aggregated active demand samples (1440 per day) during the training period (here, 3 months), $P_{1+1440} \ldots P_{n+1440}$ are the active demand samples for the day ahead relative to $P_1 \ldots P_n$, and T, WS, H and DT are temperature, wind speed, relative humidity and day type for the corresponding periods. Total aggregated reactive demand (Q) was forecast in the same way. This approach is based on the load forecasting methodology described in [16].

$$Input = \begin{pmatrix} P_1 & \dots & P_n \\ T_1 & \dots & T_n \\ WS_1 & \dots & WS_n \\ H_1 & \dots & H_n \\ T_{1+1440} & \dots & T_{n+1440} \\ WS_{1+1440} & \dots & WS_{n+1440} \\ H_{1+1440} & \dots & H_{n+1440} \\ DT_1 & \dots & DT_n \\ DT_{1+1440} & \dots & DT_{n+1440} \end{pmatrix} \qquad (3.14)$$

$$Target = [P_{1+1440} \quad \dots \quad P_{n+1440}] \qquad (3.15)$$

The results of the active and reactive load forecast are shown in Figure 3.30.

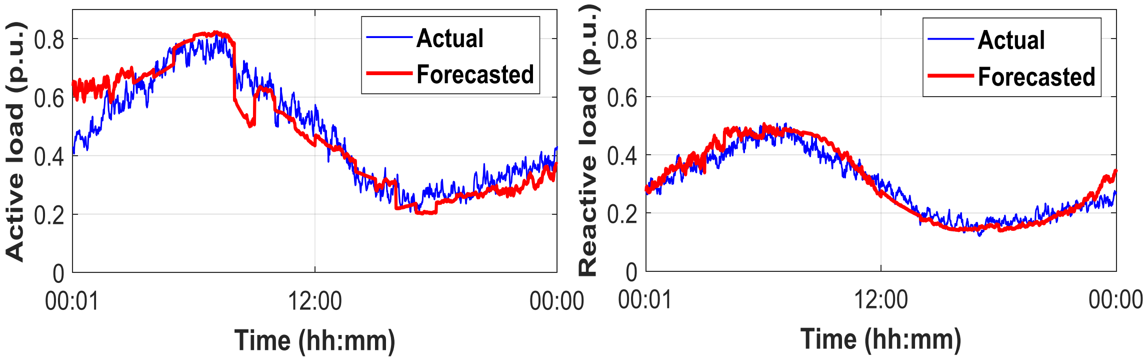
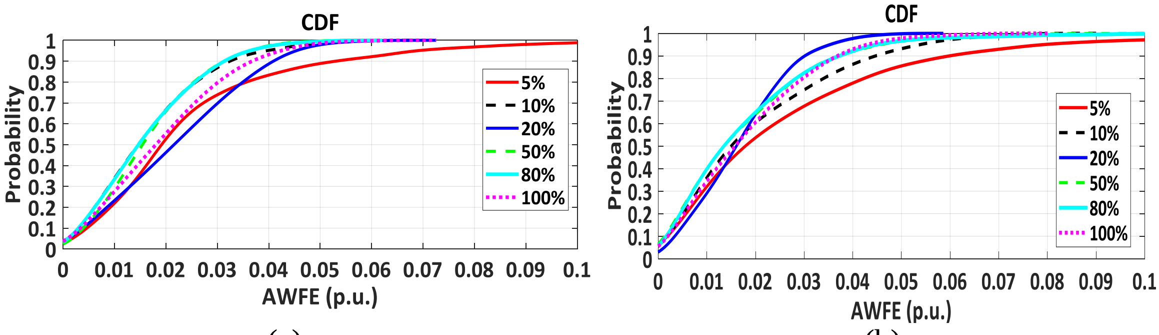


Figure 3.30 Day ahead total load forecast for August 30th 2015

As the mean of the absolute percentage error ($MAPE$, defined in (3.16), where n is the number of samples, x is the calculated value, and x_0 is the actual value) is 11% for active load and 9% for reactive load, a demand decomposition test was done by incorporating a 10% $MAPE$ for both P and Q in the ANN input. Therefore, two scenarios were examined for the ANN input: $0.9 * P$ and $0.9 * Q$ as the lower bound input ($MAPE = -10\%$), and $1.1 * P$ and $1.1 * Q$ as the higher bound input ($MAPE = +10\%$). The resulting distribution of AWFEs for controllable active demand forecasting is presented in Figure 3.31 for -10% total load forecasting error (Figure 3.31a) and +10% total load forecasting error (Figure 3.31b). When compared to the same case presented in Figure 3.28 (left), which is based on the accurate values of total P and Q, it can be seen that there is no degradation in accuracy.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{x_i - x_{0,i}}{x_{0,i}} \right| \qquad (3.16)$$



(a) (b)

Figure 3.31 Controllable load estimation for one day in August with 7 days training (August), with (a) day-ahead total load forecasting error of -10% and (b) day-ahead total load forecasting error of +10%

Finally, Figure 3.32 illustrates the controllable load forecast with 5% SM coverage based on the accurate total active and reactive load forecast (thick line), and the range of values considering ±10% MAPE in the forecast of total active and reactive load. The actual controllable load curve is given by the dotted red line. As seen in the figure, the differences are typically very small and only exceptionally approaching 8-10%. This leads to the conclusion that reasonable total load forecasting errors do not drastically affect the accuracy of load decomposition. For reasons of clarity, only a part of the day is presented in the figure.

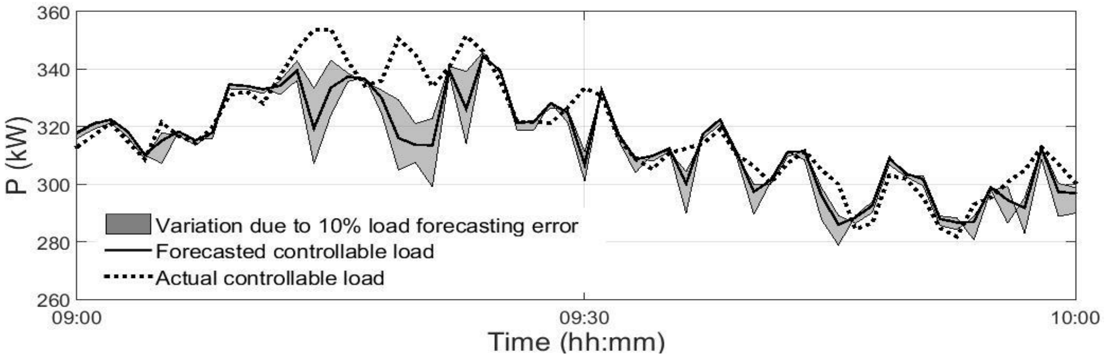


Figure 3.32 Forecast and actual controllable load with the possible variation range due to the 10% error of the total demand forecasting

3.7 Comparison of ANN and ARIMA method for demand decomposition

This section tests the eligibility of ARIMA method [201] for demand composition forecasting, as it has been widely used in the past for time series forecasting purposes [202]. If all the end-users in the observed aggregation have sub-metering facilities, based on the historical daily curve of each individual load category, it would be possible to forecast (in the short-term, e.g., day ahead) each category's demand. However, in cases of limited number of end-users sending their sub-metering data, ARIMA output would have to be scaled up to the total number of users. In order to assess how this could affect the accuracy of demand decomposition, testing was performed with the ARIMA model, and the influence of SM coverage level on the accuracy of results was investigated. The forecast of the load composition, namely the amount of controllable load, was done for day-ahead based on the most recent 7 days historical data. The ARIMA model was implemented in Matlab [203].

For illustration and comparison purposes, the loading curve of controllable load (denoted as *y*) was taken from the data set described in Section 3.4 (aggregation of 1000 users whose demand was generated using CREST model [111]). It was assumed, as previously, that there were missing samples in the SM data, which were processed using LI method. The data had to be additionally pre-processed, as it was non-stationary, i.e., there was a high autocorrelation observed over the entire time series. The non-stationarity can be seen from the autocorrelation function (ACF) of the time series *y* (with minute-based samples) in Figure 3.33, where the correlation coefficients are far outside the significance limits (defined by parallel blue horizontal lines) at almost all lags (time steps). Thus the data had to be differenced (consecutive samples in the time series were subtracted from each other) before further processing [204]. As the differencing of minute-based data made no improvements, the time series was first smoothed (through averaging over 30 min periods), and then differenced. Finally, the new time series (with 48 samples per day) had improved stationarity, as the ACF and partial autocorrelation function (PACF) show in Figure 3.34 and Figure 3.35, respectively. Correlation index at most lags stayed around the significance limits.

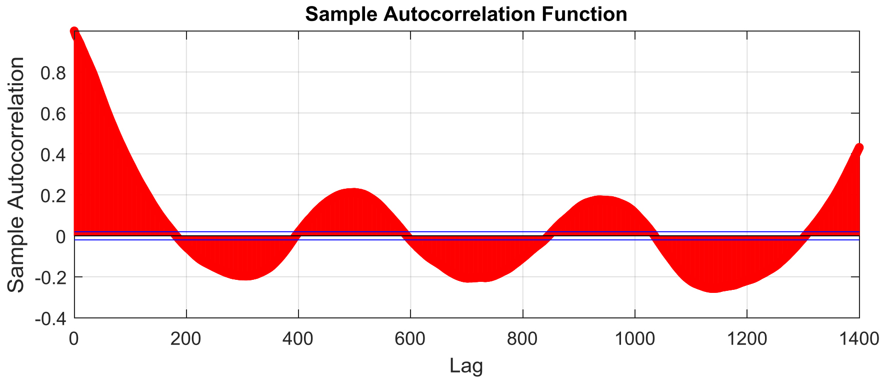


Figure 3.33 Autocorrelation function of the original samples *y*

The next step was to estimate the parameters of the ARIMA model, which is commonly done based on the ACF and PACF of the time series. The parameters are: non-seasonal auto-regressive term (p), non-seasonal differencing term (d), non-seasonal moving average term (q), seasonal auto-regressive term (P), seasonal differencing term (D), seasonal moving average term (Q) and seasonality (S) [203]. Based on ACF and PACF of the differenced data, the chosen ARIMA $(p, d, q)(P, D, Q)S$ model was $(1,1,1)(1,1,1)48$. These values were adopted considering the fact that the seasonality was 48 samples (daily), and that the most significant time steps in ACF and PACF were around time step 1, while the time series was differenced once.

For the consistence needed for comparison with the ANN approach, the time series forecasting (day ahead) of the controllable load was performed based on the last 7 days historical values of the data. The outputs obtained for different SM coverage levels were appropriately scaled in order to be compared to the base case (aggregation of 1000 users). For example, the output (demand of controllable load during one day) obtained with 20% SM coverage level was scaled by factor 5 before it was compared with the corresponding values in the base case. Figure 3.36 and Figure 3.37 represent an example of controllable load forecast based on the measurements from end-users at different SM coverage levels (5-20% in Figure 3.36 and 50-100% in Figure 3.37). The dotted lines refer to the actual values for days 7 and 8, while the solid lines refer to the ARIMA forecast for day 8, at the corresponding SM coverage level.

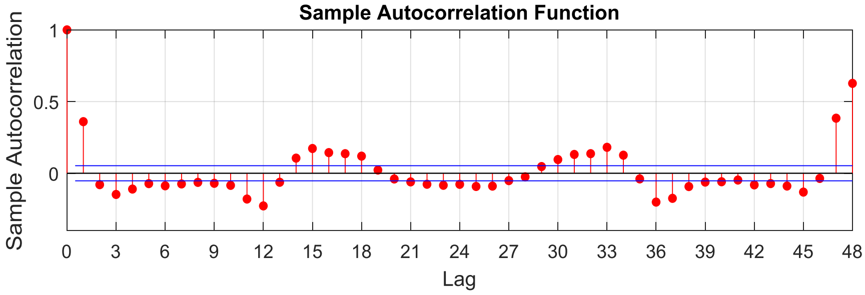


Figure 3.34 Autocorrelation function of the differenced samples of y

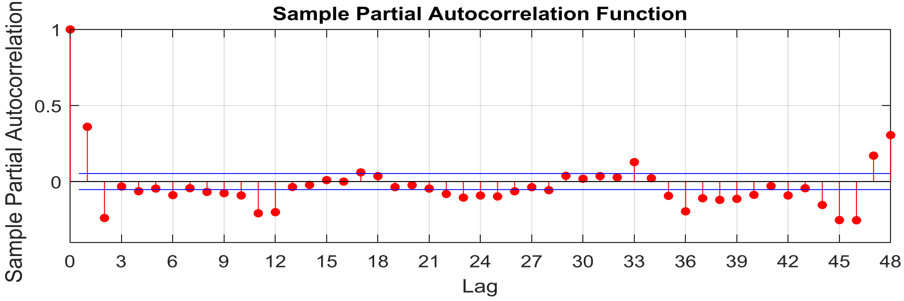


Figure 3.35 Partial autocorrelation function of the differenced samples of y

Finally, if the obtained forecast of the controllable load at different SM coverage levels (5%÷100%) is scaled to 1000 users (by factors 20÷1, respectively), the results for the day 8 forecast are as shown in Figure 3.38 and zoomed in Figure 3.39.

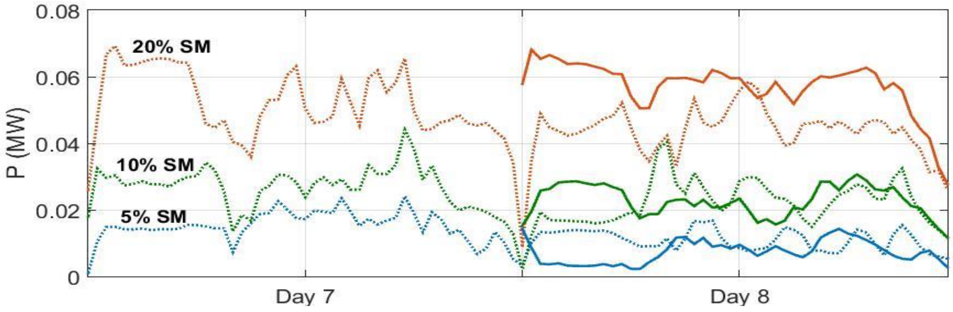


Figure 3.36 ARIMA forecast of the 8th day controllable load at different SM coverage (5-20%), with days 7 and 8 shown

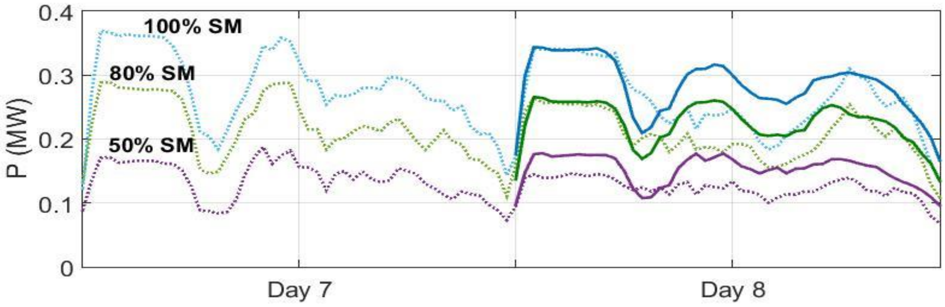


Figure 3.37 ARIMA forecast of the 8th day controllable load at different SM coverage (50-100%), with days 7 and 8 shown

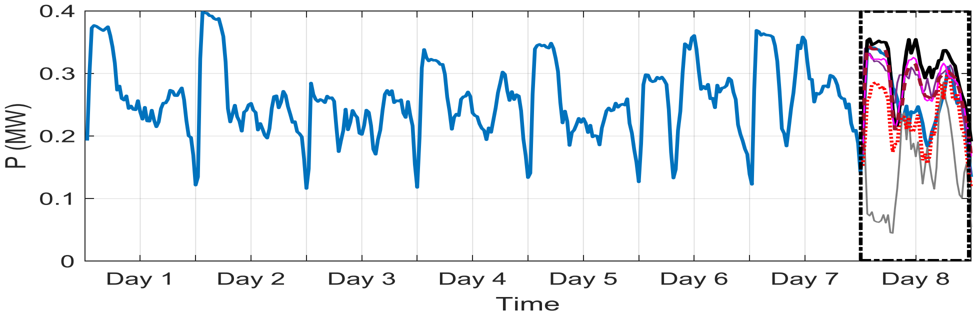


Figure 3.38 Actual load during the 7 days preceding the 8th day and the forecast based on different SM coverage

Figure 3.38 shows the 7 days data (controllable load only) used for training consisting of 7*48 samples. The observed inaccuracy (higher than in case of ANN), even at 100% SM coverage, can be related to high variability in daily loading curve of the controllable load, as well as fewer points used for training. In case of ANN, 7*1440 samples were used.

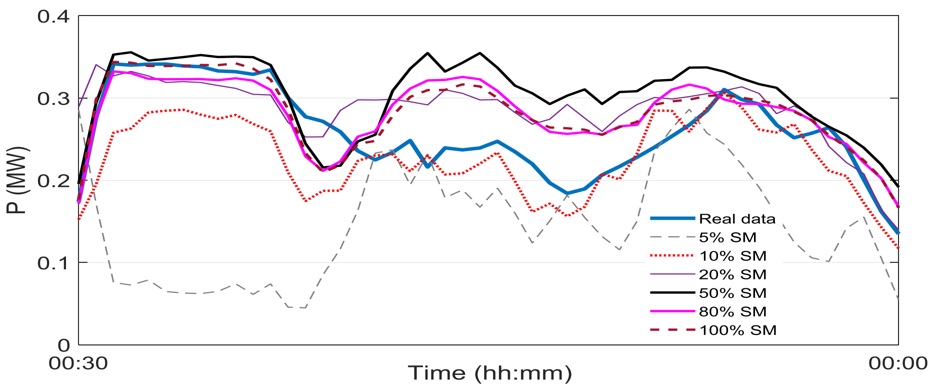


Figure 3.39 ARIMA forecast of the 8th day consumption of controllable load based on scaled outputs at different SM coverage

The AWFE was further calculated for simulation results for the whole month of August (this way, the number of output samples was 30*48=1440, which corresponded to the number of output samples in the cases described in Section 3.5.1) and shown in Figure 3.40. The results were obtained iteratively, by running the ARIMA model for one day based on the most recent 7 days and repeating this process for every day in the month. It can be seen from the figure that even at higher SM coverage levels, the errors obtained with ARIMA are visibly larger than with the ANN based methodology (see Figure 3.41, which is a reproduction of Figure 3.15). The 90th percentile of the error ranges between 0.15 p.u. and 0.3 p.u., depending on the SM coverage level. It should be noted that in the ANN approach, these errors ranged from 0.05 p.u. to 0.1 p.u. (as shown in Figure 3.41). This brings the conclusion that at 5% SM coverage, the 90th percentile AWFE using ARIMA method is 3 times larger than with ANN, hence using ARIMA method for demand composition forecasting is not justifiable considering more accurate ANN based forecasting.

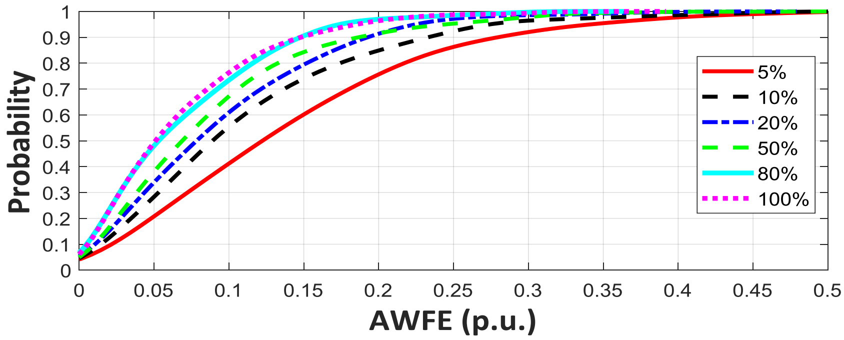


Figure 3.40 CDF of AWFE for controllable load during one month

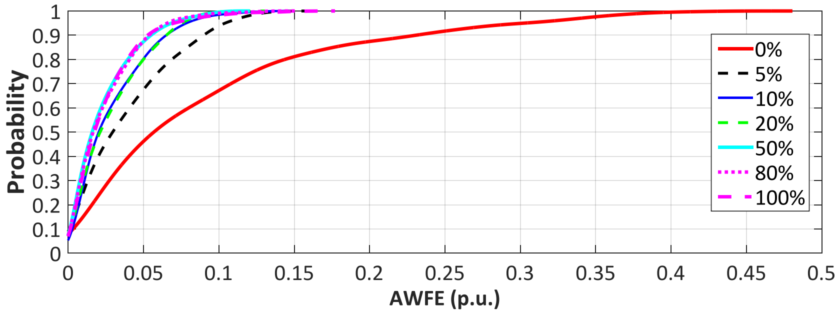


Figure 3.41 Reproduction of Figure 3.15 CDF of AWFE for the estimation of controllable active load, case C

3.8 Graphical user interface for representing demand composition

The graphical user interface (GUI) for advanced demand profiling (ADP) of the residential load is foreseen as a decision making tool for a DR responsible party (e.g., an aggregator or DNO). The overall ADP methodology has been initiated within the UK EPSRC project “Autonomic power systems” and EU FP7 SUSTAINABLE project and fully developed within the EU H2020 NOBEL GRID project as a micro-service of the integrated NOBEL GRID Consumer Profiling Framework. It incorporates two functionalities: mid-term (day ahead) and short term (half hour ahead) demand forecasting (DF) and demand decomposition (DD), the latter relying on the former. The reason for this interdependence is that the necessary input for DD is total active and reactive demand, as described in Section 3.4. As most of the DR programs are planned based on the forecast flexibility of the end-users (e.g., in [69]), the information about demand composition should also be given in advance, most commonly one day ahead. Therefore, as a necessary step before DD, information about forecast active and reactive demand has to be obtained first. The forecast is done at the aggregation point, which is usually the substation supplying the consumers participating in the DR programme. Other scenarios, where the aggregation involves consumers scattered around a geographical (and electrical network) area, are also possible.

DF is therefore done at the aggregation level, either at the substation or for a group of scattered customers belonging to the same aggregator, and do not necessarily have to be connected to the same LV substation. The forecast is performed either a day or half hour ahead and it has two main applications:

- distribution network operation, where the forecast load is used as the necessary information for balancing between demand and available generation, and

- input for the DD module which uses the total demand prediction to forecast demand composition.

The former application could be complemented by a renewable energy sources (RES) forecast, which allows low carbon operation of the distribution network, while the latter one actively supports the DR programs, by providing information on forecast amount and type of controllable loads of the end-users.

The information flow between the two modules of the ADP and the GUI is illustrated in Figure 3.42. The input data (namely historical and real-time demand and weather data) are first sent to the DF module, whose results (forecast active and reactive demand) are then fed to the DD module, for finally obtaining the composition of the forecast demand.

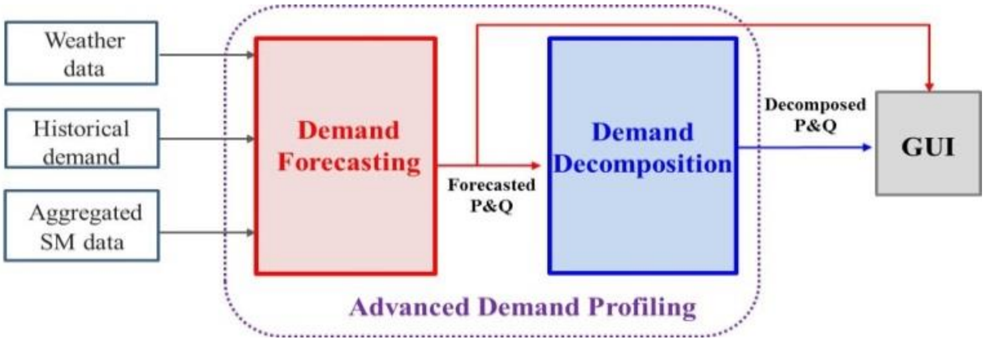


Figure 3.42 Flowchart of the information flow

The demand forecasting module is based on the use of ANN, following methodology for load forecasting introduced in [86], which uses historical demand data (active/reactive load) and weather data (temperature, humidity and wind speed) for training the ANN, and most recent historical demand and forecast weather data for obtaining the half hour or day ahead demand forecast. The details of this approach were provided in Section 3.6.2. The training of the ANN is presented in Figure 3.43, where the three rows for weather data correspond to temperature, humidity and wind speed, and day type refers to working/non-working day value (1/0). Once the ANN is trained, the input data should be in the same format as the training inputs. In order to retain acceptable accuracy of the DF module, historical data is regularly updated, with the fixed time window of the training data. This means that the ANN for DF (for simplicity, referred here as DFANN) is always trained with the same size of historical data (in this example, 3 months) by dismissing the oldest samples every time the training data is updated with new measurements, for example daily or weekly.

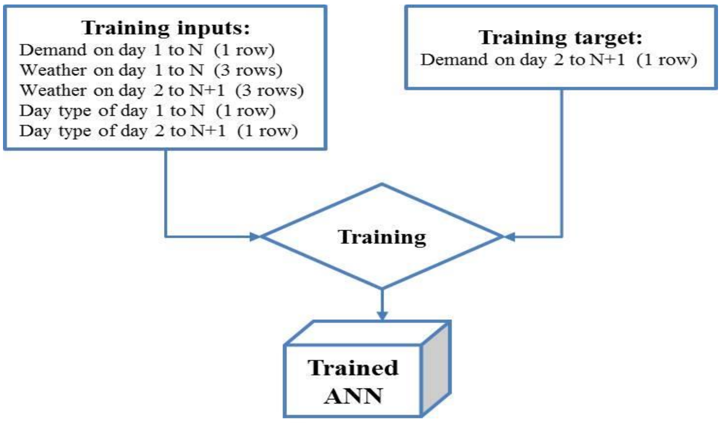

Training inputs:
Demand on day 1 to N (1 row)
Weather on day 1 to N (3 rows)
Weather on day 2 to N+1 (3 rows)
Day type of day 1 to N (1 row)
Day type of day 2 to N+1 (1 row)
Training target:
Demand on day 2 to N+1 (1 row)
Training
Trained ANN

Figure 3.43 Training process of the DFANN

Besides historical demand and the weather data necessary for training the DFANN, the update rate of weather forecast (input) data is also important. The DF module was tested using test data from an actual test site (substation), where historical data measurements of active demand (with half hour resolution) and corresponding weather (one hour resolution) were available. The weather forecast for the next 24 hours would be updated once/twice a day, which affected the accuracy of results. Figure 3.44 illustrates the importance of input data update on the accuracy of day ahead DF. Datasets 1 and 2 represent forecasting outputs where weather data is updated every 12 hours and 24 hours, respectively. The relative errors (in %) based on the forecast and actual (measured) load values are given over a 36 hour period (72 time steps).

Between time steps 14 (corresponding to 7 a.m.) and 38 (corresponding to 7 p.m.), as well as between steps 61 and 72, dataset 1 is forecast based on the updated weather data, while dataset 2 still uses the old weather forecast (which only gets updated at 7 p.m.), resulting in lower accuracy, as highlighted in the dotted frames in Figure 3.44. After this period, during the next 24 steps (12 hours), both datasets are forecast based on the same weather data, which is why the errors are the same during that period. In order to show the dependence on the weather update, Table 3.7 presents the most probable relative errors (MPRE) for day ahead and half hour ahead forecasting, for datasets 1 and 2. As seen in the table, the weather update shows a higher impact on the accuracy of day ahead forecasting than on half hour ahead forecasting. It should be noted that two different DFANNs are generated for day ahead and half hour ahead forecasts. The reactive demand forecast can be analysed in a similar way, as long as the reactive power measurements (or power factor values, as in this test site's data) are available. In the next step, demand decomposition is performed following the methodology described in Section 3.4.

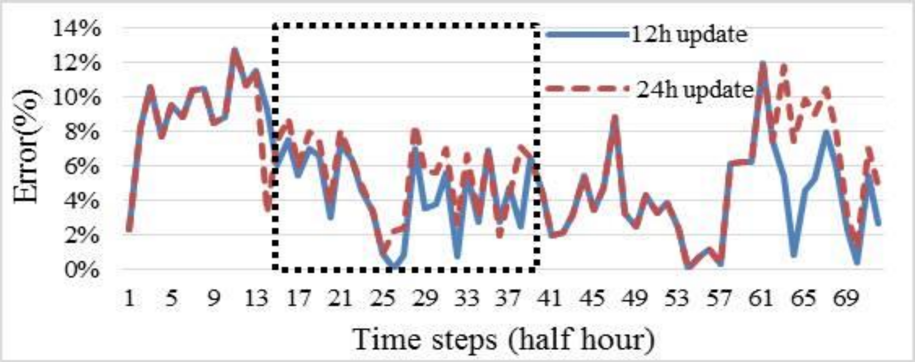


Figure 3.44 Relative error of demand forecasting for datasets 1 and 2 over 36 hours (72 time steps)

Table 3.7 The most probable relative error for different demand forecasting cases

Case	Dataset 1		Dataset 2	
	Day ahead	Half hour ahead	Day ahead	Half hour ahead
MPRE	**4.71%**	2.38%	**6.74%**	2.40%

Although DF is an on-line application, there is a certain time delay between the input data acquisition and the output – forecast active/reactive load. Similarly, additional delay will be introduced with DD module, as the ANN for DD (denoted as DDANN) relies on the output of DF and takes some time for the simulations. The delays will depend on the data latency of the communication system responsible for gathering the measurements, data pre-processing time and also on the computational power (processor). The overall process of ADP, fully developed within the NOBEL GRID project, is illustrated in Figure 3.45. The input data (weather, historical demand and real-time SM data) are fed into the ADP engine, consisting of DF and DD modules, each introducing a certain time delay (Δt_1 and Δt_2, respectively). With the 64-bit operating system and 3.40 GHz processor used in this example, delay introduced by the DF module was 5 seconds, while the delay caused by the DD module was 7 seconds, giving 12 seconds of total delay for the overall ADP process. The results are transmitted to the Demand Response Flexibility Market (DRFM) Cockpit. The DRFM Cockpit was foreseen in the NOBEL GRID project as an intermediary system between the aggregated flexibility of the end-users and other smart distribution grid actors (e.g., the DSO or the aggregator).

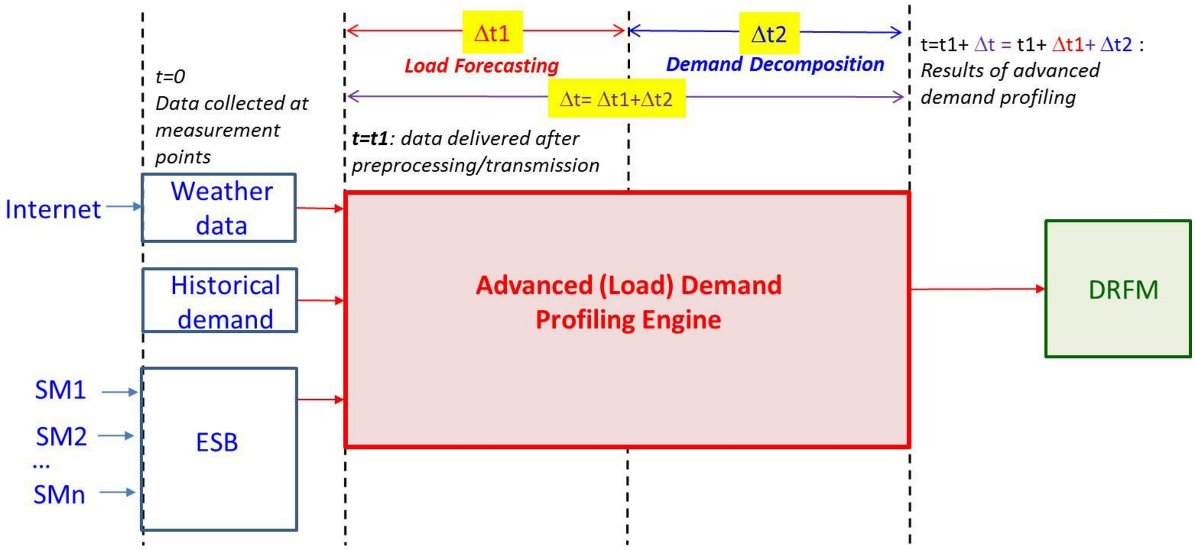

Δt1
Load Forecasting
Δt2
Demand Decomposition
Δt= Δt1+Δt2
t=t1+ Δt = t1+ Δt1+ Δt2 :
Results of advanced demand profiling
t=0
Data collected at measurement points
t=t1: data delivered after preprocessing/transmission
Internet
Weather data
Historical demand
SM1
SM2
...
SMn
ESB
Advanced (Load) Demand Profiling Engine
DRFM

Figure 3.45 Advanced demand profiling, as seen in NOBEL GRID project

In order to present the ADP output in a user-friendly way, two GUIs have been developed, for DF and DD modules. Print-screens of the two GUIs are presented in Figure 3.46 and Figure 3.47, showing the ADP results using the historical total active/reactive demand and weather data from the aforementioned test site. As the site had no sub-metering data, the training data for DDANN was generated using Monte Carlo simulations and statistical data about the load composition in residential areas, similarly to the approach described in [39] and mentioned in Section 3.5.1. Both GUIs (for the DF and DD module) offer day ahead and half hour-ahead forecasting. As the GUIs are updated with new measurements arriving in real-time, the actual total demand (from substation measurements) and actual demand composition (resulting from the DDANN taking measured P and Q data as input), are presented, enabling real-time comparison and error calculation between the forecast and actual values. The two GUIs "communicate" with each other, the output of the DF module being used as the input for DD module.

Figure 3.46 represents the GUI for DF, where different numbered parts of the interface have the following function/meaning:

- 0: Activate/deactivate button serves to start or interrupt the GUI. It should be noted that the DFANN is first trained with the most recent, in this case 3 months-long, historical weather and demand data. Once trained, only the last half hour/24 hour demand and half hour/24 hour ahead weather forecast data is needed as the input to perform the half hour/24 hour ahead demand forecast;

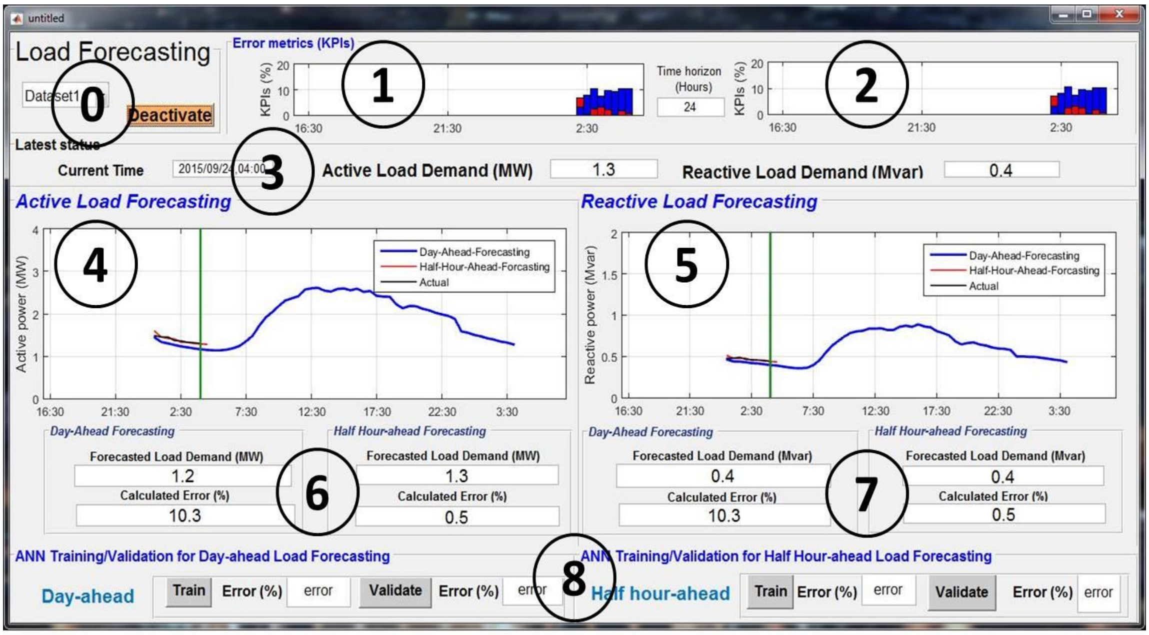
- 1 and 2: Bar plots showing the time change of the key performance indicators (KPIs), i.e., relative errors between the forecast and actual values for real and reactive demand, respectively. Red bars refer to the half-hour ahead forecasting errors, and blue ones to the day-ahead forecasting errors, which are usually higher. The presented KPIs are illustrative - additional indicators, for example mean absolute percentage error (MAPE) [86], can be added;

- 3: The ‘Latest status’ shows the current time (here, 24/09/2015 at 04:00), and current (real-time) measured active and reactive demand at the substation;

- 4 and 5: Graphs presenting day ahead (in blue) and half hour-ahead (in red) forecast active and reactive demand, respectively, together with the actual (measured) demand (in black). The vertical green line refers to the current time – as the time changes, the demand curves move to the left, while the green line stays fixed. The curves will move with the same time resolution as the resolution of the measurement data, with some time delay, as mentioned. Therefore, the curves on the right side of the green line refer to the forecast demand, while those on the left side correspond to the most recent historical data, measured and forecast;

- 6 and 7: Numerical values of day-ahead and half hour-ahead forecast active and reactive demand, respectively, which were forecast for the current time, followed by the calculated relative error based on the actual (measured) demand at the current time;

- 8: Training and validation of the ANN for day-ahead and half hour-ahead forecasting, respectively. There is an optional “Train” button for the user, who may want to retrain the DFANN with the most recent (3 months long) historical data, while the ‘Validate’ button performs validation of the ANN, i.e., testing the ANN with training data and comparing the outputs with the ANN target. The retraining can also be done automatically, after a pre-defined time period, e.g., every day or weekly.



Current Time 2015/09/24 04:00

Active Load Demand (MW) 1.3

Reactive Load Demand (Mvar) 0.4

Day-Ahead Forecasting: Forecasted Load Demand (MW) 1.2; Calculated Error (%) 10.3

Half Hour-ahead Forecasting: Forecasted Load Demand (MW) 1.3; Calculated Error (%) 0.5

Day-Ahead Forecasting: Forecasted Load Demand (Mvar) 0.4; Calculated Error (%) 10.3

Half Hour-ahead Forecasting: Forecasted Load Demand (Mvar) 0.4; Calculated Error (%) 0.5

Day-ahead: Train Error (%) error; Validate Error (%) error

Half hour-ahead: Train Error (%) error; Validate Error (%) error

Figure 3.46 Presentation of the GUI for day ahead and half hour ahead demand forecasting

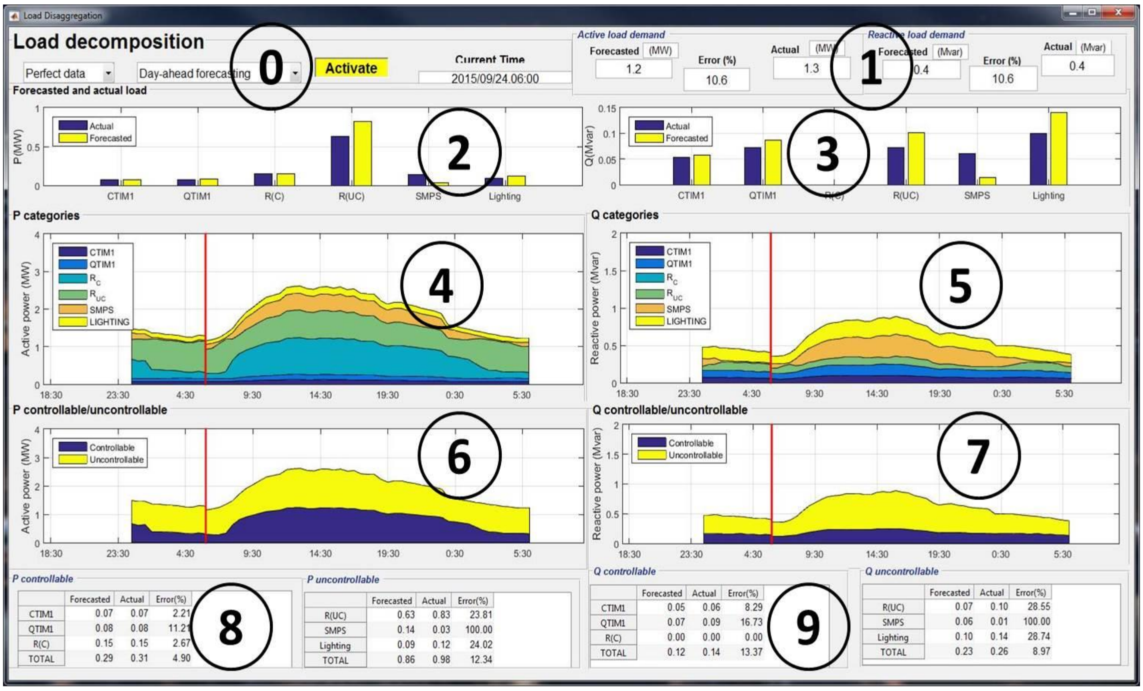
Figure 3.47 represents the GUI for DD, where numbered parts of the interface have the following function/meaning:

- 0: The drop-down menus offer different datasets (in this case “perfect data”, i.e., data with no missing samples was chosen) and half hour or day-ahead forecasting. The Activate/Deactivate button serves to start or interrupt the GUI. Current time is also shown here.

- 1: Values of relative errors for active and reactive total demand forecasting (showing numerical values of the DF outputs for the current time);

- 2 and 3: Bar plots showing the forecast (yellow) and actual (blue) demand of individual load categories, within total real and reactive demand, respectively;

- 4 and 5: Diagrams showing forecast demand composition based on the forecast total active and reactive demand (illustrated in Figure 3.46) on the right side of the red vertical line, and actual demand composition, based on measured total active and reactive demand, on the left side of the line. The red line refers to the current time. Similarly to the case of DF, the curves move in real time to the left, together with the time labels, while the red line stays fixed. Also, the curves will move with the same time resolution as the resolution of the measurement data, with some time delay.



Perfect data | Day-ahead forecasting

Current Time: 2015/09/24.06:00

Active load demand: Forecasted (MW) 1.2; Error (%) 10.6; Actual (MW) 1.3

Reactive load demand: Forecasted (Mvar) 0.4; Error (%) 10.6; Actual (Mvar) 0.4

P controllable

	Forecasted	Actual	Error(%)
CTIM1	0.07	0.07	2.21
QTIM1	0.08	0.08	11.21
R(C)	0.15	0.15	2.67
TOTAL	0.29	0.31	4.90

P uncontrollable

	Forecasted	Actual	Error(%)
R(UC)	0.63	0.83	23.81
SMPS	0.14	0.03	100.00
Lighting	0.09	0.12	24.02
TOTAL	0.86	0.98	12.34

Q controllable

	Forecasted	Actual	Error(%)
CTIM1	0.05	0.06	8.29
QTIM1	0.07	0.09	16.73
R(C)	0.00	0.00	0.00
TOTAL	0.12	0.14	13.37

Q uncontrollable

	Forecasted	Actual	Error(%)
R(UC)	0.07	0.10	28.55
SMPS	0.06	0.01	100.00
Lighting	0.10	0.14	28.74
TOTAL	0.23	0.26	8.97

Figure 3.47 Presentation of the GUI for day ahead forecasting of demand composition

- 6 and 7: Diagrams showing forecast controllable and uncontrollable load (on the right hand side of the vertical red line) based on the forecast total active and reactive demand, and historical-actual controllable and uncontrollable load, based on measured total active and reactive demand. These curves move accordingly.

- 8 and 9: The values represent the forecast and actual active and reactive demand, respectively, of individual load categories, and controllable/uncontrollable load. These are followed by the corresponding relative errors.

The main purpose of advanced demand profiling and the GUI is to support day ahead and half hour ahead DR planning, as it provides information about the flexibility potential of the demand-side over different times of the day. This information reduces the uncertainty of the actual (available) flexibility of the end-users, even before a DR signal is sent to them by the DR responsible party. The variety of information provided by the GUI can be further adjusted based on the specific requirements of the user. Furthermore, information about the types of flexible load (motors, heaters, lighting, etc.) that could be potentially shifted to different time (disconnected) enables a proper network performance analysis before any kind of load shifting or load curtailment is performed.

The introduced graphical representation of demand-side flexibility at the local (substation or other aggregating point) level can be further extended to a whole network and presented in “geographical map format” showing, in both time and space, the varying flexibility of the load at each individual substation (bus) in the network at any given point in time. The developed GUI for demand forecasting and demand decomposition, as a means of representing demand size and its flexibility in a user friendly way, can be used in a control room by the distribution system operator, who will be able to forecast, with high confidence, when, where (at which buses) and how much the demand can participate in, and increase the flexibility of, the low carbon network daily operation.

3.9 Summary

This chapter presented a methodology for aggregated demand decomposition using limited SM data and the application of ANN. The overall methodology results in an estimation of the shares of different load categories and controllable and non-controllable load within the total forecast demand, with a foreseen application in various DR programs. The proposed methodology can be used for either real-time or forecasting applications, although the focus of this chapter was on day-ahead forecasting. The two main assumptions are that the SMs can record active load of individual appliances and that only some of the end-users in the aggregation have this type of SM. Special attention has been given to data pre-processing, i.e., aggregation of data streams coming from SMs in different sampling steps and with missing samples. Two data restoration methods were analysed in this respect, linear interpolation and weight adjusted k-nearest neighbour method. An approach for obtaining probabilistic aggregated reactive load curve is also discussed as a solution to the lack of reactive load measurements at the end-users’ point. The methodology was further validated on an actual pilot site’s dataset, resulting in similar accuracy and confirming that there is no need for weather data and large historical data to be included. Since the consumption data is aggregated at a data concentrator point (the substation), the size of the data sent to the upstream network, e.g., to the DNO or an aggregator, is also reduced. In addition, the ANN-based approach was compared to ARIMA method, showing its superiority. Development of the methodology for aggregated demand decomposition using limited SM data represents the third original contribution of this thesis.

In order to fully develop and implement the demand decomposition methodology described above it was necessary to develop an approach for estimating reactive daily loading curve. The methodology is established based on probabilistic modelling of PF of individual appliances and subsequent Monte Carlo simulations. This methodology represents the fourth original contribution of the thesis.

Finally, a GUI for advanced demand profiling was presented and discussed in this chapter. It is a practical tool foreseen to be used in a distribution network control centre for short-term DR planning. Development of the graphical user interface for advanced demand profiling represents the fifth original contribution of the thesis.

4 Multi-objective Demand Side Management at Distribution Network Level

4.1 Introduction

This chapter illustrates the concept of multi-objective DSM in a distribution network in support of transmission network operation. The methodology builds on the results of the methodology on Advanced Demand Profiling, detailed in the previous chapter. Information about demand composition is used to model demand at each load bus of the network, facilitating that way further studies of the effect DSM may have on network performance indicators. Multi-objective DSM optimises sequentially demand scheduling, i.e., shifting of demand from one time to another, by taking into account three objectives: *i)* meeting the predefined loading curve at GSP, *ii)* preservation of demand composition, *iii)* maintenance/improvement of distribution network loadability. The optimisation takes into account, sequentially, realistic constraints of the DSM programs, namely limited demand flexibility and load payback. The proposed optimisation algorithm is foreseen as a decision making tool used by the DNO, as part of the day-ahead planning of the DSM (load scheduling) program to meet the requirements at the GSP specified by the TSO, while keeping the selected (one or more) network performance indicators within predefined limits. The concept can be equally applied to transmission network level, and to other types of network performance indicators.

4.2 System performance indicators and load modelling

In addition to providing services to the TSO, e.g., providing desired demand profile at the GSP over a set period of time, the distribution network should maintain the standard of its own network performance before and after the DSM action. This requirement is particularly important when a large number of users are involved in a DSM program, as this could substantially change the power flows in the distribution network, and such influence, to an extent, the transmission network operation as well. Different network performance indicators can be observed, individually or in combination, such as frequency, voltage levels, line flows, network losses, etc. These indicators will be, more or less, affected by large-scale DSM depending on the network architecture (e.g., a microgrid in connected or isolated mode) and voltage levels (for example, frequency is more critical in transmission network than in distribution network).

As already detailed in the introductory chapter, Section 1.5.2.4.1, voltage stability margin in distribution network, namely network loadability, is chosen as a network performance indicator observed before and after a DSM action. The main reason for this is the proliferation of DER and large residential loads (EVs and heat pumps) in the distribution network and possible effect this could leave on voltage profiles and loadability of both, distribution and transmission network.

Multi-objective DSM observes demand side from the two aspects:

1) The effect of time-changing composition of demand on demand flexibility and the corresponding load model parameters at each load bus in the distribution network;
2) The effect the changes in demand size and composition (i.e., load model parameters) following a DSM action can have on the distribution network performance.

The composite (ZIP+IM) load model, as one of the most comprehensive load models [45, 48], is used in this methodology in order to account for different voltage dependency of different load types, both static and dynamic, and the effect it may have on system steady state and dynamic performance. Static performance is observed through load flows and loadability assessment, while the dynamic performance is observed through the dynamic response of demand. Load at every load bus in the network is therefore represented using the ZIP+IM model.

In order to illustrate dependence of network loadability (load margin) on the load model, PV curves were generated in DIgSILENT/PowerFactory on an example of a simple 2-bus network representing a distribution network with a GSP, as given in Figure 4.1. Three cases were observed: constant power load model, ZIP load model and ZIP+IM model. In all three cases the total network load was equal to 0.8 MW, with reactive load equal to 0.16 Mvar. The r/x ratio of the distribution line equals 1.2, while the equivalent model of the network represents the upstream network.

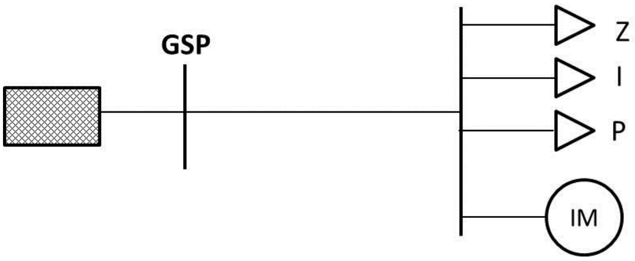


Figure 4.1 Simple network model for PV curve simulation

Results in Figure 4.2 show that constant power model gives the most conservative critical load (corresponding to the tip of the PV curve in Figure 1.4), in this case 2.9 MW, while ZIP model gives considerably larger critical load equal to 4.6 MW, due to the existence of voltage sensitive load components (namely constant current and constant power load model). Finally, ZIP+IM model results in slightly lower load margin, showing 4.4 MW as the critical loading of the system. It can be concluded that if a more realistic load model is used, such as the ZIP or ZIP+IM load model, the critical loading is higher, and consequently, the load margin, i.e., the distance between the current operating point and the critical loading is larger. Even though the load margins resulting from the ZIP and ZIP+IM load model are not very different, the reason IM load is included in the load model is to capture the dynamic response of composite demand, which occurs following a disturbance in the network (e.g., voltage step change), and depends on the size of demand and demand composition. Furthermore, considering that many of the load components in the network that could participate in DSM are based on IMs, the use of the composite load model is appropriate to reflect the demand composition mix before and after DSM.

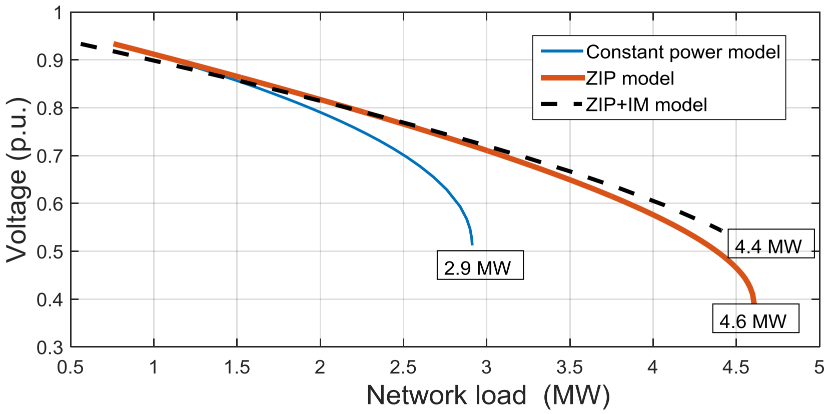


Figure 4.2 PV curve with different load models

Simplified distribution network, given in Figure 4.3, is used to show the dependence of dynamic response of demand on the composition of demand. The equivalent model of the network is used to represent the upstream network.

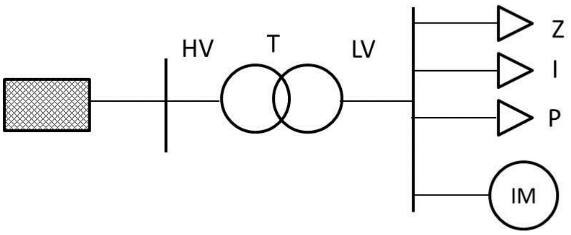


Figure 4.3 Simple network model with for dynamic response of demand simulation

Three cases are investigated to illustrate the effect of different shares of ZIP and IM loads in the total demand, namely 80:20, 50:50 and 20:80, respectively. Figure 4.4 represents dynamic response of demand following a 5% voltage drop due to a transformer tap change on transformer T shown in Figure 4.3.

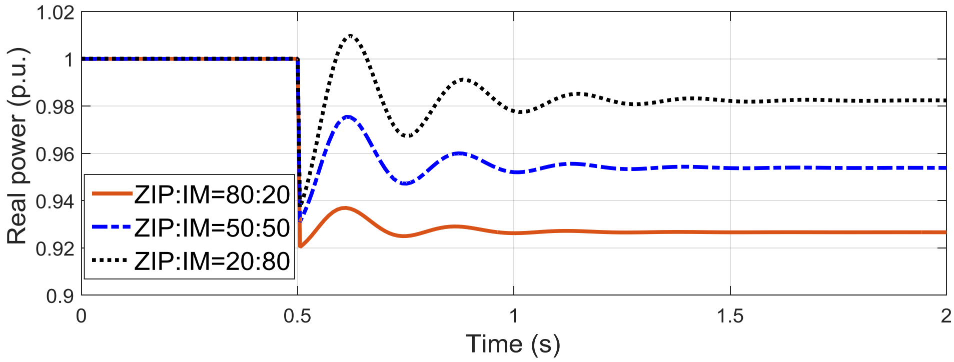


Figure 4.4 Dynamic response of demand following a disturbance due to 5% voltage drop (tap change)

The responses confirm that the higher the share of dynamic (IM) loads is, the more oscillatory is the real power response of demand to a voltage disturbance, and larger steady state value following a disturbance is attained. The system disturbance was simulated on a 33/11 kV transformer, while the response was recorded on the HV side of transformer. The total nominal network demand was 40 MW. This is an illustrative example only of the extent to which demand composition may influence the dynamic response of demand.

4.2.1 Load categories and composite load model

The six load categories identified in Chapter 3 are further grouped into the four components of the composite load model (constant Z loads, constant I loads, constant P loads and IM loads). Mapping between these two types of demand classification is given in Table 4.1. An example of a daily loading curve decomposed into load categories and composite load model components is given in Figure 4.5. Furthermore, Z loads are divided into controllable and uncontrollable based on their suitability for DR, while all IM loads are considered as controllable. The rest of the components are deemed uncontrollable. The composite load model at each load bus will thus consist of the 4 components (ZIP+IM), whose shares in the total load are different across different load buses, and also change during the day. Each load component comprises one or more load categories, which were originally defined in [86] and presented in Section 3.2.

Table 4.1 ZIP+IM Load Components and Corresponding Load Categories

ZIP+IM model component	Load category
Z_C (controllable constant impedance loads)	R_C
Z_{UC} (uncontrollable constant impedance loads)	R_{UC}
I (uncontrollable constant current loads)	Lighting
P (uncontrollable constant power loads)	SMPS
IM (controllable induction motors)	CTIM1+QTIM1

The proposed demand composition may be simplistic from a point of view of an aggregator, as individual appliances have different behaviour depending on operating cycles (for example, a washing machine may be operating predominantly as a motor or as a heater, depending on the part of the operating cycle). These operating features are of great importance to the entity controlling and scheduling demand. From the network operator’s point of view, however, it is deemed unnecessary to observe these differences among different appliances. The methodology mainly focuses on aggregated demand, where volatility of operating cycles of individual loads is less visible. The aim of demand composition presented in this chapter is to enable as realistic as possible modelling of demand using the most advanced, composite (ZIP+IM) load model, and to illustrate the extent to which demand flexibility can be harnessed if realistic load modelling is used to represent demand, and intrinsic demand limitations (with respect to flexibility and load payback) are taken into account.

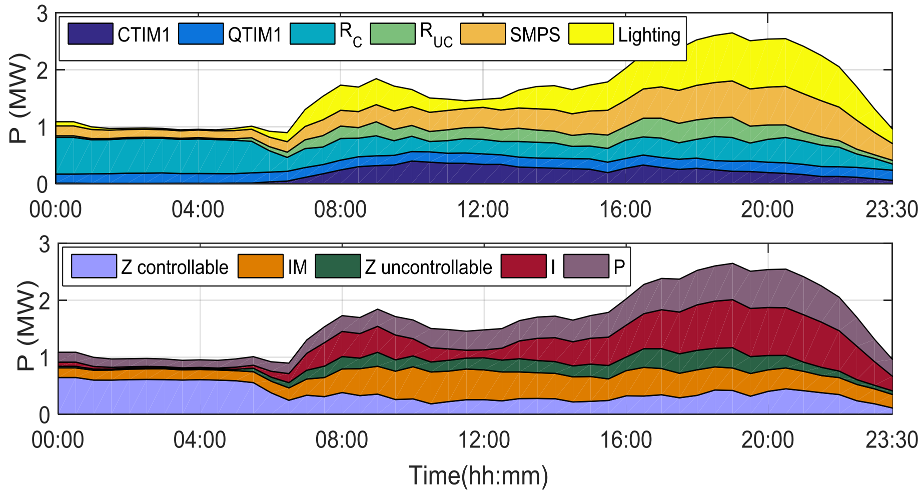


Figure 4.5 Demand composition of active demand: 6 load categories (top) and ZIP+IM composition (bottom)

4.2.2 Load payback

Load payback, i.e., the reconnection of previously disconnected loads at the time steps following the disconnection, is accounted for in this methodology in order to illustrate a realistic DSM program. One of the most common ways to model load payback is using linear increase of load [205], represented by the following equation:

$$PB(t) = \alpha \cdot \Delta(t-1) + \beta \cdot \Delta(t-2) + \gamma \cdot \Delta(t-3) \qquad (4.1)$$

where $PB(t)$ is the payback load at time step t, Δ is the amount of shifted load (in MW, for example), and α, β and γ are the payback coefficients for load shifted from the three preceding time steps (three hours). This linear model is adopted for modelling the payback of controllable Z loads in this methodology. The approach assumes that all the loads disconnected at one time step get gradually re-connected in the following three time steps, i.e., $\alpha + \beta + \gamma = 1$. It is also assumed that the approximately equal shares of the disconnected load are re-connected within the 3 hours after the DSM action, as suggested in [206]. Thus, 34% of the disconnected Z loads will be reconnected in the first hour, the following 34% in the second hour and the last 32% in the third hour after load disconnection.

In the case of IM loads, it is assumed that the users/aggregators are given the freedom to choose when to reconnect the loads, as long as it is during the valley periods (mostly night time). During the valley period, the load is lower than the one requested by the network operator. Therefore, all the disconnected IMs are reconnected randomly (following a uniform distribution), within the given periods of time and within 24 hours. A drawback of this generic approach is that IM loads also involve motors modelled as QTIM1 loads, i.e., cooling devices. These devices may not be deferrable for many hours. For example, it has been reported in [207] that food can only stay safe if the fridge has been turned off for up to 4 hours. Different reconnection schemes and their effects can be addressed in the future or in more specific case studies considering actual implementation of DSM in real networks. In this particular case study though, QTIM1 loads do not represent a significant share in total load, which is why their load payback is not modelled separately.

These two payback models account for the end-users' commodity - they consider the usage of IMs as less comfort-constrained, i.e., the end-users are more tolerant to postponing the operation of wet appliances than the operation of the heating devices. Similar availability of these load types was reported in [54]. It is important to note that appropriate communication and control infrastructure are assumed to exist in the distribution network, allowing for scheduling of the disconnection and reconnection of the load, as described in this section.

4.3 Methodology

The aim of DR, i.e., load modulation adopted in this methodology as a balancing service provided by the DNO to the TSO [101], is to flatten the daily loading curve of the distribution network by shifting some or all of the controllable loads from valley periods. Load flattening reduces the need for generator ramping, and the number of system balancing actions [54]. Any other predefined shape of the daily loading curve could have been equally adopted to illustrate the approach, for example, reducing the system peak load. As already mentioned in the introductory chapter, balancing services in National Grid are provided by aggregators in a form of reserves, by increasing, decreasing or shifting demand [109].

DSM in this methodology is planned day ahead, with three objectives, met in three consecutive steps: 1) ensuring that the distribution network load follows a predefined load profile, taking into account load flexibility and load payback; 2) preservation of demand composition; 3) preservation/improvement of the load margin.

The first objective is met by applying optimal power flow (OPF) with participation of flexible load buses acting as distributed generators with negative output. The algorithm takes into account that different buses have different flexibility (controllability) during the day, and so the load shift at any bus and time step is limited by the available (predicted) load flexibility. The output of this optimisation step informs the operator *how much of the flexible load* should be shifted at each load bus and when, to ensure that the set load profile is followed as closely as possible.

The second objective is to keep the demand composition after DSM as close as possible to the one before DSM using linear programming. As previously mentioned, demand composition, given as the contribution of individual load components (e.g., induction motors, resistive loads, etc.), plays an important part in the dynamic response of demand following a disturbance in the network that could ultimately lead to voltage and/or angular instability [86]. Maintaining same/similar demand composition after a DSM action reduces the possibility of unexpected load behaviour in case of a disturbance. The output of this optimisation step therefore informs the operator *what portion of a particular type of flexible load* should be shifted at each bus and when.

Finally, at the third step, the load margin at each time step (of the 24-hour planning horizon) is checked and compared with the one before DSM. If at any time step of the planning horizon the load margin after DSM appears to be lower than a pre-specified tolerance range around the load margin before the DSM, the load dispatch (shift) is corrected at the corresponding time step. In this methodology, for illustration purposes, the pre-specified tolerance range is chosen to be 5%, i.e., the new load margin should not be lower than 95% of the initial load margin. Other tolerance values can be used equally, depending on the criteria chosen by the network operator. Similar approach for the load margin tolerance level was reported in [76], where the critical network load was 5% lower than the network loading corresponding to the tip of the PV curve (maximum network loading). Furthermore, in contingency analysis transmission system operators define a minimum loading margin to ensure that the current operating point has a minimum distance to the collapse point [208]. For each contingency, the system has to ensure a minimum loading margin. If a contingency leads to load margin lower than this one, then it is considered critical and requires corrective actions. Otherwise, if a contingency is characterised by maximum loading level lower than the current operating condition, that contingency is unfeasible.

The correction of the voltage stability index (VSI), i.e., the load margin, is applied by optimising load values of the controllable load components (Z and IM loads) using Particle Swarm Optimisation (PSO). The PSO method was chosen due to its proven applicability in economic dispatch [209] and OPF including voltage stability indicators [210], as well as higher computational speed over genetic algorithms [211], another heuristic optimisation method used in similar problems.

Figure 4.6 illustrates the main steps of the proposed methodology. Once the loads are dispatched at every load bus of the network using OPF calculations (Level 1), load margin is assessed (Level 2) by running the PV curve simulations. If voltage stability index (VSI) is lower than the threshold, the PSO is used to re-schedule the flexible loads within their flexibility boundaries (lower and upper bound, *lb* and *ub*, respectively) to allow for higher load margin (Level 3). The steps are further detailed in the following sections.

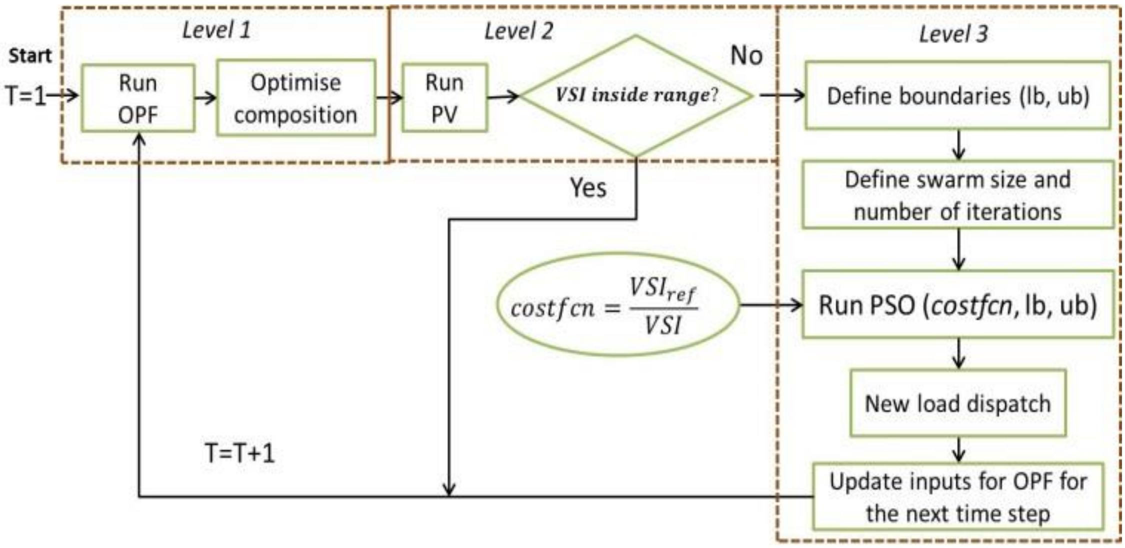

Level 1
Level 2
Level 3
Start
T=1
Run OPF
Optimise composition
Run PV
VSI inside range?
No
Yes
Define boundaries (lb, ub)
Define swarm size and number of iterations
$costfcn = \frac{VSI_{ref}}{VSI}$
Run PSO (costfcn, lb, ub)
New load dispatch
Update inputs for OPF for the next time step
T=T+1

Figure 4.6 Flowchart of the methodology

4.3.1 Optimal power flow

By flattening the load curve and minimising the load flow through the GSP, the distribution network acts as a balancing service provider. The dispatch of flexible demand to meet, in an aggregated way, the desired loading curve, can be represented as an optimisation problem. The problem is solved as a typical OPF, given by expression (4.2), where the cost of generation and load shift is minimised. In this case, the highest generation cost is assigned to the GSP, while the flexible loads (acting as DGs with negative output) have the lowest cost in order to ensure that load follows generation. The generation cost of the DG only formally participates in cost minimisation, as the output of the DG is fixed and serves only to model the target loading curve that aggregated flexible loads should adjust to. The cost function of the DG, as well as the cost function of the slack bus (which is in OPF equivalent to a generator), are given in Table A3 in the Appendix A. It should be noted that the generation cost itself is irrelevant, as long as the supply from the GSP is costly enough to allow for demand flexibility to be prioritised over the GSP, i.e., the upstream network supply when solving the OPF. Generation cost function is quadratic for the GSP, to minimise the flow through the slack bus.

$$min\left(\sum_{j=1}^{N_G} C_j P_{G,j} + \sum_{j=1}^{N_D} C_j P_{Disonnected,j}\right) \tag{4.2}$$

Subject to:

$$P_{G,i} - P_{D,i} = V_i \sum_{k=1}^{N} V_k [G_{ik} \cos\theta_{ik} + B_{ik} \sin\theta_{ik}] \tag{4.3}$$

$$Q_{G,i} - Q_{D,i} = V_i \sum_{k=1}^{N} V_k [G_{ik} \sin\theta_{ik} - B_{ik} \cos\theta_{ik}] \tag{4.4}$$

$$V_i^{MIN} \leq V_i \leq V_i^{MAX} \tag{4.5}$$

$$P_{D\,i,t} = P_{Forecasted,i,t} - P_{Disconnected,i,t} + P_{Connected,i,t} \tag{4.6}$$

$$0 \leq P_{Disconnected,i,t} \leq \Delta_{i,t} \tag{4.7}$$

(4.3) and (4.4) are power flow equality constraints; (4.5) refers to bus voltage limits; (4.6) takes into account that the load value at each bus and each time step depends on the size of shifted (disconnected) load (calculated by OPF), and the size of payback (re-connected) load at that time step; (4.7) presents the flexibility limits of the load, where Δ is the amount of flexible load.

OPF is run in Matpower [212], which allows modelling of flexible loads as generators with negative output. The software uses constant power model only. In order to validate usage of two types of software, namely Matpower and DIgSILENT/PowerFactory in the analyses described in this chapter, comparison of power flow results obtained with them is given in the Appendix A. For comparison purposes, load flow in DIgSILENT/PowerFactory was run using the constant power load model.

4.3.2 Preservation of demand composition

In this step, demand composition, i.e., the shares of the four load components of the composite load model (ZIP+IM) at every bus, is kept as close as possible to the one before DSM. This will ensure that the demand at GSP will also have similar composition before and after DSM, at the given season and time of the day, hence similar dynamic response following network disturbance. This is achieved by maximising the sum of shares of shifted (disconnected) IM loads ($\mathrm{P^{IM}}_{Dis.}$) and Z loads ($\mathrm{P^{Z}}_{Dis.}$), as shown by (4.8), while at the same time constraining these shares to ensure that demand composition is preserved. The constraints are given by (4.9) which preserves the ratio of controllable loads, defined by ΔZ and ΔIM, before (left hand side of the equation) and after DSM (right hand side), and by (4.10) which keeps the disconnected loads within corresponding flexibility limits. $P^{IM}{}_{Con.}$ and $P^{Z}{}_{Con.}$ in (4.9) are payback (reconnected) IM and Z load, respectively, at the corresponding load bus and time step. (4.11) limits the sum of the disconnected shares by the total disconnected load calculated at the first optimisation level, i.e., the maximum disconnected load cannot exceed the value determined by OPF.

$$max(\mathrm{P^{Z}}_{Dis.} + \mathrm{P^{IM}}_{Dis.}) \tag{4.8}$$

Subject to:

$$\frac{\Delta IM}{\Delta Z} = \frac{\Delta IM - P^{IM}{}_{Dis.} + P^{IM}{}_{Con.}}{\Delta Z - P^{Z}{}_{Dis.} + P^{Z}{}_{Con.}} \tag{4.9}$$

$$0 \le P^{Z}{}_{Dis.} \le \Delta Z \ ; \ 0 \le P^{IM}{}_{Dis.} \le \Delta IM \tag{4.10}$$

$$\mathrm{P^{Z}}_{Dis.} + \mathrm{P^{IM}}_{Dis.} \le P_{Disconnected} \tag{4.11}$$

This optimisation step is performed in Matlab. If the OPF tool could incorporate the composite load model, levels 1 and 2 of the optimisation could be merged and solved together in Matpower, although this process would not be trivial from the modelling perspective. Therefore, due to the practical limitations of the software in use, two-level optimisation is used to solve OPF and optimal load composition problem.

4.3.3 Particle swarm optimisation

PSO belongs to the group of heuristic optimisation methods, along with genetic algorithms and evolutionary algorithms [213]. These methods start from a random choice in the search space and, based on the evaluation of the objective function at every iteration, gradually move the position of the result vector to the optimal one. In the PSO method, a swarm (population) of candidate solutions (particles) is generated in the first iteration, and the positions of particles are updated in the following iterations based on the values of the objective function. The basic PSO algorithm consists of three steps: generating positions and velocities of the particles, velocity update, and position update [211]. The initial positions and velocities are allocated randomly, from the search space, and the velocities are updated in the following iteration based on the values of the fitness function of the particles within a swarm. The velocity update ($\mathrm{V_i^{t+1}}$) uses information about the particle with the best global value in the current swarm (the so called local best – $P_{i_{best}}^{t}$), and the best position of any particle over time (the so called global best solution – G_{best}^{t}). Finally, the particle position is updated based on the velocity update. The position of the i-th component of the particle vector X (X_i^{t+1}) is updated based on the previous time step t, and following (4.12) and (4.13):

$$X_i^{t+1} = X_i^t + V_i^{t+1} \tag{4.12}$$

$$V_i^{t+1} = c_1 V_i^t + c_2 rand(0,1)(P_{i_{best}}^t - X_i^t) + c_3 rand(0,1)(G_{best}^t - X_i^t) \tag{4.13}$$

where c_1, c_2 and c_3 are acceleration constants, defining the linear attraction towards the direction of the particle. Coefficient c_1 defines the tendency of the particle to continue in the same direction, while c_2 and c_3 define attraction towards the local best (found by the given particle at any iteration) and global best solution (found by any particle at any iteration), respectively [214]. The first coefficient should not be too large or too small, to prevent slow or premature convergence, respectively. The extensive studies reported in [215] showed that the optimal value for c_1 is 0.7 or 0.8, while the value for c_2 and c_3 is between 1.5 and 1.7.

4.3.4 Load margin – based PSO algorithm

The PSO algorithm is used to reschedule the controllable load shift every time the load margin after DSM is estimated to be lower than 95% of the load margin before DSM. As already mentioned in Section 1.5.2.4.1, load margin reflects the distance of the current operating point of the system to the maximum loading point, and is commonly determined from the active load–voltage characteristic (the PV curve), as shown in Figure 1.4. Keeping the load margin as large as possible ensures that the system is able to withstand increase in demand or disturbances without endangering its voltage stability. In order to ensure that the load margin is maintained, the aim of the PSO algorithm is to minimise the objective function (4.14) subject to (4.10), where VSI_{ref} is the load margin prior to the DSM action, and VSI is the load margin after rescheduling the demand shift of IM and controllable Z loads.

$$min\left(\frac{VSI_{ref}}{VSI}\right) \tag{4.14}$$

The PSO algorithm is applied in Matlab, with the swarm size (number of particles) set to 100. Matlab default values for acceleration constants of $0.1 \leq c_1 \leq 1.1$, and $c_2 = c_3 = 1.5$ are used. Even though [215] recommended a fixed value of 0.7 or 0.8 for c_1, the PSO algorithm in Matlab uses adaptive value from the given range ($0.1 \leq c_1 \leq 1.1$) during the iterations. Additional simulations were run to compare these two approaches, and the results showed that fixed value (0.7 was used in this case) in some cases reduces the simulation time (around 15 minutes instead of around 22 minutes when using adaptive c_1), however it results in notably higher cost function (lower values of the load margin) in some time steps. In other cases, fixed value of c_1 results in slightly lower cost function (higher values of the load margin) than with adaptive value, but increases simulation time from 19 to 23 minutes. Therefore, the adaptive value of c_1 was used eventually. The cost function of the applied PSO algorithm converged to a fixed value on average after 13 iterations, hence this number was chosen as the maximum number of iterations in order to reduce the computational time. This number was obtained by observing the convergence process (an example is shown in Figure 4.7). The simulations were run with a larger number of iterations and the cost function value was recorded at every iteration. Once this value stabilised over a number of iterations, the required number of iterations was fixed. It takes up to around 22 minutes for the overall algorithm to run (for 24 hour planning horizon, i.e., 48 time steps), including the PSO simulations with 13 iterations, on a PC with the 64-bit operating system and 3.40 GHz processor. There is certainly a scope for reducing the computational time, however, this is not affecting the implementation of the proposed methodology, hence it was not deemed necessary to be considered as a part of this stage of the research.

The stopping criterion of the iterative process is determined by either the maximum number of iterations (13 in this case) being reached, or when the cost function reaches a limit value equal to 1 (i.e., the new load margin is equal to the one before DSM). Therefore, the aim of the optimisation is not necessarily to maximise the load margin, but to keep it unchanged after the DSM action. The man reason for this is reduction of computational time. It should be noted that the minimum cost function of the PSO algorithm is not obtained when all resources of one or both controllable load types are disconnected, but rather when a certain combination of the two, determined by the algorithm, is reached.

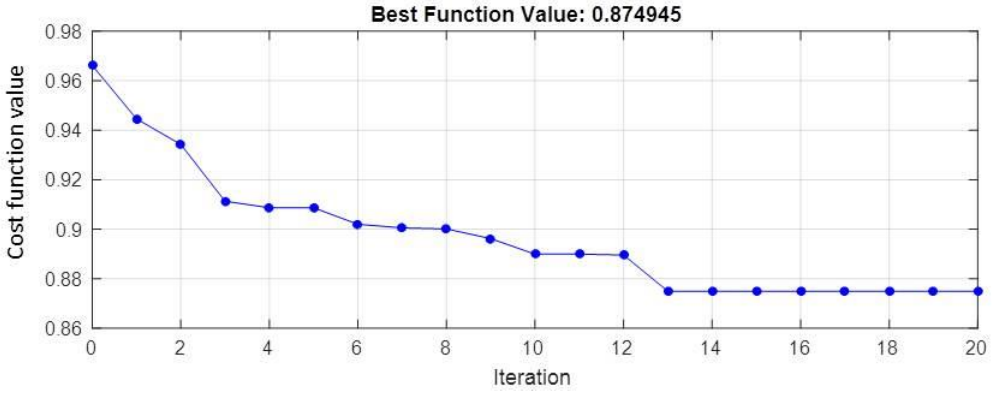


Figure 4.7 Cost function evolution over 20 iterations

Further illustration of the methodology is given in Figure 4.8 to clarify the steps of the PSO algorithm, and distinguish between parts of the methodology realised using different pieces of software (Matlab, Matpower or DIgSILENT/PowerFactory). As already mentioned, OPF is run in Matpower, resulting in real and reactive load values at each load bus (P_{load} and Q_{load}). Optimisation of demand composition is then performed in Matlab, resulting in optimal demand values of controllable load groups, namely IM loads and controllable Z loads (P_{IM} and P_Z in Figure 4.8). In the next step, PV curve simulations are run in DIgSILENT/PowerFactory, using the composite load model with P_{IM} and P_Z values. If the obtained load margin is lower than the threshold, PSO algorithm is applied, starting with initialisation of $N = 100$ particles in the swarm. Each particle is a vector containing relevant demand values of IM and controllable Z loads at all the load buses in the network (32 of them), assigned randomly by the algorithm, respecting the load flexibility limits at each load bus, as given by (4.10). The cost function of the PSO algorithm (given by (4.14)) is calculated based on the outputs of the PV simulations performed in DIgSILENT/PowerFactory for every particle in the swarm. The iterative process updates the swarm following (4.12-4.13) until the stopping criterion is met. Finally, based on the new optimal values of P_{IM} and $P_{Z,}$ load values are updated (together with the load payback in the upcoming time steps following the resulting load shift in the current time step) and the overall process moves to the next time step of the planning horizon.

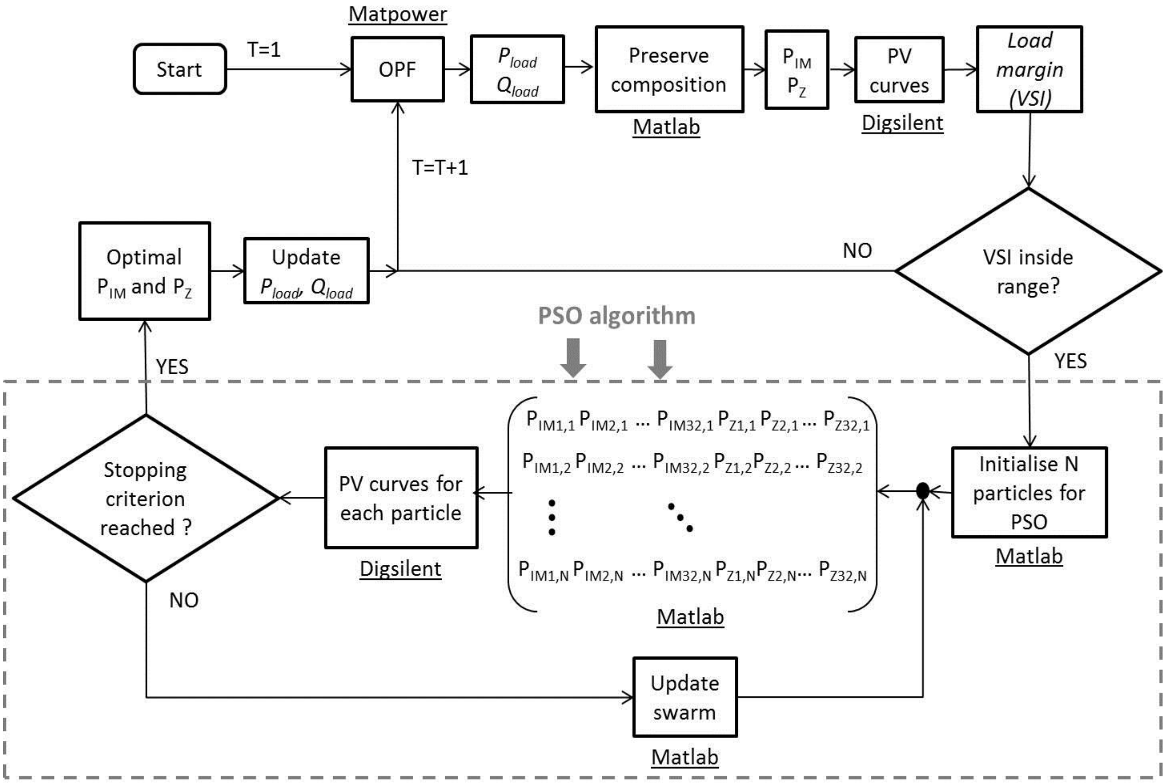

Matpower
Start
T=1
OPF
P_load Q_load
Preserve composition
Matlab
P_IM P_Z
PV curves
Digsilent
Load margin (VSI)
T=T+1
Optimal P_IM and P_Z
Update P_load, Q_load
NO
VSI inside range?
PSO algorithm
YES
YES
Stopping criterion reached ?
PV curves for each particle
Digsilent
P_IM1,1 P_IM2,1 ... P_IM32,1 P_Z1,1 P_Z2,1 ... P_Z32,1
P_IM1,2 P_IM2,2 ... P_IM32,2 P_Z1,2 P_Z2,2 ... P_Z32,2
P_IM1,N P_IM2,N ... P_IM32,N P_Z1,N P_Z2,N ... P_Z32,N
Matlab
Initialise N particles for PSO
Matlab
NO
Update swarm
Matlab

Figure 4.8 Flowchart of the methodology with detailed steps of the PSO algorithm

As the overall algorithm relies on network simulations run on two types of software, namely DIgSILENT/PowerFactory and Matpower, validation of the IEEE 33-bus distribution network model (the original model already exists in Matpower) in DIgSILENT/PowerFactory was performed by running a power flow on the network model with loads modelled as constant power. The results were identical to the results of power flow in Matpower, which validated the network model (see Table A4 and Figure A1 in Appendix A). This proved the validity of transferring results from one software environment to the other.

4.4 Case studies

The test network used in this study is a slightly modified IEEE 33-bus distribution network shown in Figure 4.9. The network has one GSP, modelled as a slack bus, and one DG (network parameters are given in Appendix A). It represents a distribution network in two possible operating scenarios:

1) Distribution network (DN) providing ancillary services to the transmission network by reducing the need for balancing operations (by keeping the load at GSP as flat as possible, or as requested by the TSO);
2) Close to self-sufficient DN, relying on power generated by its distributed generation, and adjusting its flexible load to locally generated power.

These two scenarios will be analysed in the case studies to follow. The DG has a constant output, and the loads are dispatched in order to follow the available generation and minimise the flow through the slack bus. The two aforementioned network operating scenarios can be presented using this DG: *i)* if the network operates as a distribution network providing ancillary services to the TSO, the DG simulates the arbitrary load profile set/requested by the TSO; *ii)* if the DN operates as a self-sufficient network, it minimises its dependency on the rest of the upstream network by controlling flexible loads to follow the available generation from the DG. In both scenarios, only load flexibility is harnessed, and no changes in DG output are made.

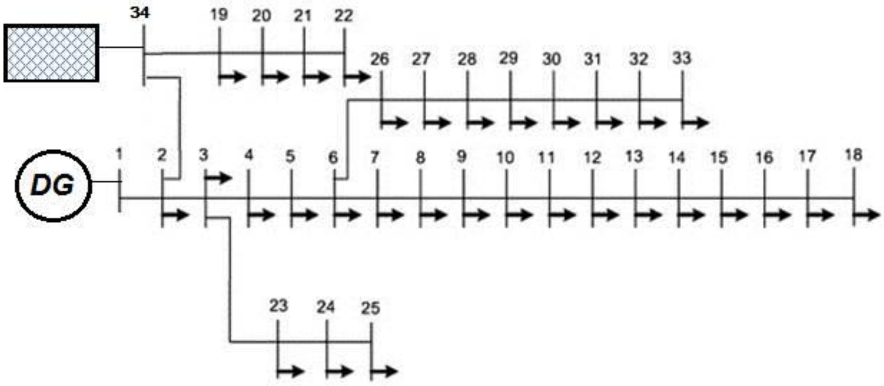


Figure 4.9 Modified IEEE 33-bus network

The network model is slightly modified compared to the original one: the added line 2-34 has the same impedance as line 1-2, while line 34-19 has the same impedance as line 19-20. All the load buses (32 of them) in the network are considered as controllable, however each load bus has different load composition and hence different controllability during the day. Each flexible load bus in the network model represents a secondary substation supplying 50 residential end-users. The default load values for the network buses (real and reactive power in the IEEE model, as given in Appendix A) are taken as the maximum daily values. The assigned (normalised) daily load curves and demand composition are generated using the CREST load model, as detailed in Section 3.5.1. The generated load curves for individual users were aggregated to 50 end-users at each load bus. For simplification, power factor (PF) is taken to be the same for all IMs and equal to 0.8 (which was the average value of the PF for the observed Pecan street dataset [24] described in Section 3.5.2). Since the uncontrollable loads do not get shifted, the only PF change in the total load will come from shifting IMs and controllable Z loads (which are considered to have unity PF). As the typical consumption of residential IMs ranges between several hundreds and several thousands of Watts [216], it is adopted, for simplicity reasons, that each IM connected to the load bus in DIgSILENT/PowerFactory has the load of 1 kW. This is required as the change in consumption of the IMs in DIgSILENT/PowerFactory is modelled by changing the number of motors (each having a constant load of 1 kW) connected in parallel at a load bus. Higher granularity than this one was deemed unnecessary.

It should be noted that DSM results in the following subsections are presented over a time period of 24 hours or 36 hours (ending at noon of the following day instead of midnight of the initial day) in order to illustrate clearly the effect of the load payback (which often happens during the low load period of the day following a day with load shift) on the outcome of the DSM program.

4.4.1 Operating scenario 1: Demand profiling based on external request (i.e., distribution network providing balancing service)

The proposed DSM methodology is demonstrated on a set of seven case studies, listed in Table 4.2. In each case the effectiveness of load shaping is evaluated with peak to average ratio (PAR) [54] – the closer this ratio is to 1, the more successful load shaping (flattening) is. In addition, distribution network losses were observed before and after the DSM action, to evaluate the extent to which load shaping contributes to their reduction, as losses represent a significant share in the overall operational costs of a DNO, affecting greenhouse gasses emission and generator capacity requirements [98]. In the case study A, a subcase with preserved composition (case A.1) was compared to the cases when composition is not preserved, but either IM (case A.2) or Z (case A.3) loads are prioritised (disconnected first) to meet the desired load shift. If all the resources of one flexible load type are used up in cases A.2 and A.3, and they still do not meet the demand reduction requirement, the appropriate amount of the other load type is then disconnected. In some time steps this may lead to disconnection of all the flexible loads (both prioritised and the other).

Table 4.2 Case studies

Case study	Subcases
A. Base case (3 MW peak load)	A.1 Preserved composition
	A.2 Prioritisation of IM loads
	A.3 Prioritisation of Z loads
B. Case with limited acceptability	B.1 Preserved composition
C. Overloaded system	C.1 Preserved composition
D. Critically loaded system	D.1 Preserved composition
E. Neglected DSM constraints	E.1 Constant power model

Case B observes a scenario where different load buses show different shares of customers which accept to participate in the DR program. Therefore, it is assumed that different buses have, randomly, 20%, 50% or 80% acceptability level, which reduces the DSM potential. Case studies C and D illustrate operating scenarios of overloaded and critically loaded systems, respectively. Based on the approach reported in [217], the power transformer is overloaded if the loading is between 1.25 and 1.5 times higher than its rated loading (kVA). Therefore, it was assumed that the network loading is 1.5 times higher than the rating of the transformer at GSP (the transformer is not represented in the network model in Figure 4.9). It can be assumed that the transformer rating is calculated as follows:

$$R = P_{max} + a \cdot S \tag{4.15}$$

where P_{max} is the peak network loading, S is the standard deviation (adopted to be 25%), and $a = 1.28$ is a coefficient corresponding to the 90% confidence level, adopted from the Gaussian probability table [217]. The peak network loading, given in this methodology as the nominal load of the test network, is 3.715 MW. Therefore, the transformer rating is $3.715 + 1.28 * 0.25 * 3.715 \approx 7.43\, kVA$. The base case load (case study A.1) is thus multiplied by factor equal to $1.5 \cdot (1 + 1.28 \cdot 0.25) \approx 2$. Critically loaded system (case D) is simulated by multiplying the base case load by factor 4.

Finally, case study E illustrates how different the DSM outcome is when the load is modelled as constant power (one of the most frequently used load models [48]), and limitations such as load payback and load margin are neglected.

4.4.1.1 A.1 Base case with preserved composition

Figure 4.10 illustrates the network loading curve over the 24-hour planning horizon (48 time steps) before and after the DSM. It can be seen that the resulting loading curve (solid black line) is in some time steps changed due to the activation of the PSO algorithm, triggered when there was a need to improve the load margin (Figure 4.11). As seen in Figure 4.11, the PSO algorithm successfully improves the load margin at the corresponding time steps, marked with circles, keeping it above the 95% limit. Even though the original load flexibility limit (equal to the sum of controllable loads, reflecting how “far down” load reduction can go), shown with violet dashed line in Figure 4.10, allowed for larger load decrease, load payback (green dashed line) prevented the resulting loading curve after DSM from flattening, i.e., from larger peak reduction. Finally, demand composition before and after DSM is shown in Figure 4.12. It can be seen that it is preserved to a large extent except during low load (valley) period, where the share of IM load increased due to scheduled load payback.

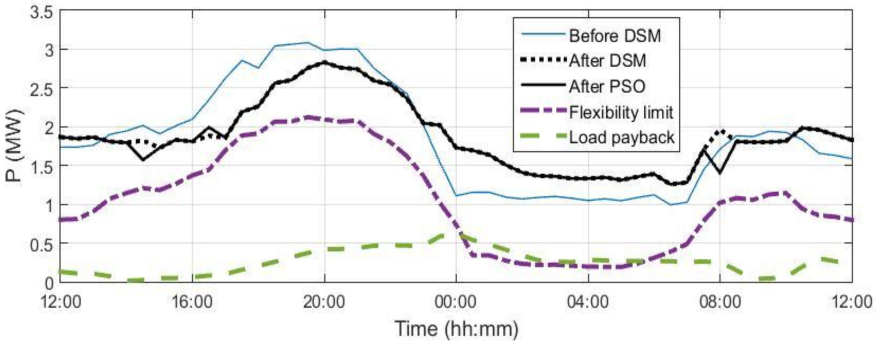


Figure 4.10 Network loading for case A.1

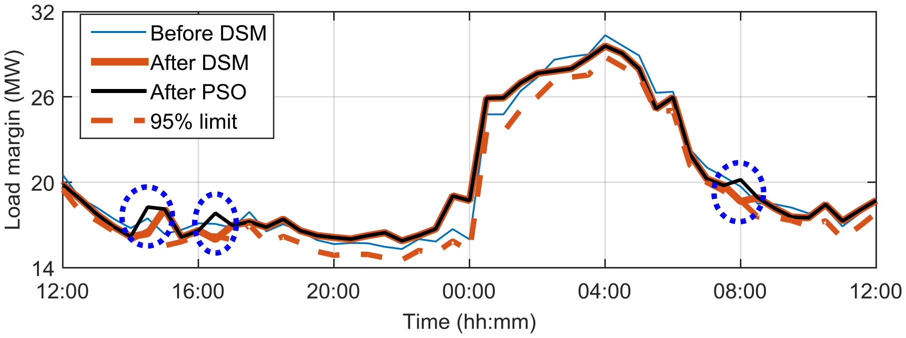


Figure 4.11 Load margin for case A.1

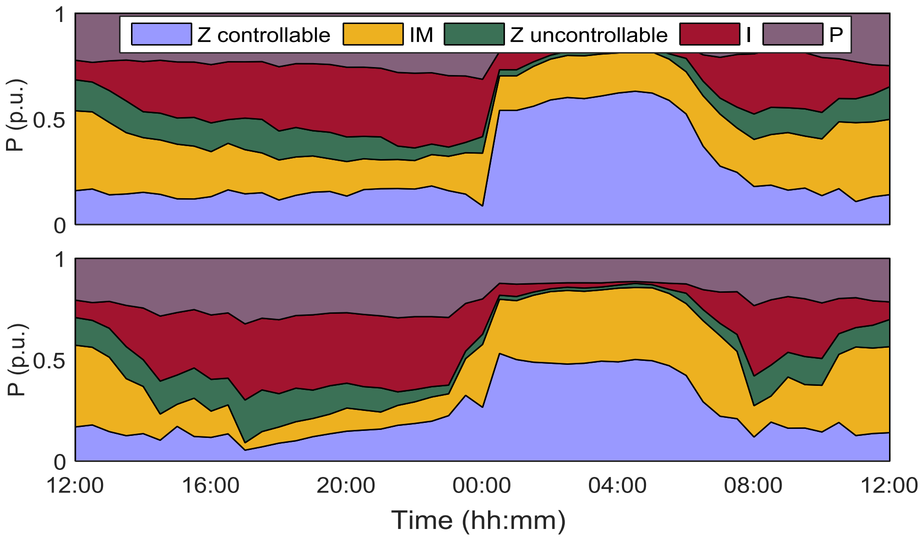


Figure 4.12 Demand composition before DSM (top) and after (bottom) for case A.1

4.4.1.2 Comparison of Case A.1 with A.2 and A.3

Figure 4.13 demonstrates how different approaches in the choice of disconnected load components can change the resulting loading curve. In these two cases the peak reduction is higher compared to case A.1 (which can also be seen from the peak to average ratio (PAR) reported in Table 4.3), however DSM action deteriorates the load margin in more time steps, which requires corrective actions (PSO algorithm), as shown in Figure 4.14, where some of these steps are marked with circles.

Table 4.3 Effectiveness of DSM in cases A.1-A.3

Subcase	Peak to average ratio (PAR) before/after DSM	Average daily losses/ Daily percentage losses before DSM	Average daily losses/ Daily percentage losses after DSM
A.1	1.66/ 1.51	0.046 MW/ 2.4%	0.043 MW/ 2.3%
A.2	1.66/ 1.44		0.039 MW/ 2.2%
A.3	1.66/ 1.40		0.040 MW/ 2.2%

Before DSM, maximum daily losses were 110 kW, while in the case of preserved composition they were 94 kW (reduced by 14%), and in case of prioritized IM or Z loads the maximum losses were 74 kW (reduced by 33%). Average daily losses and the percentage losses for the three compared subcases are given in Table 4.3. Average daily losses (P_{loss}^{avg}) and percentage losses (P_{loss}^{per}) are calculated using (4.16-4.17), where n is the number of time steps, $P_{G,i}$ and $P_{L,i}$ are network generation and network load at time step i, respectively, and $P_{GSP,i}$ is the power injected through the GSP. It should be noted that losses are calculated as the power at each time step (referring to a 30-minute period) instead of the energy.

$$P_{loss}^{avg} = \frac{1}{n}\sum_{i=1}^{n}\left(P_{G,i} + P_{GSP,i} - P_{L,i}\right) \tag{4.16}$$

$$P_{loss}^{per} = 100 \cdot \frac{\sum_{i=1}^{n}(P_{G,i}+P_{GSP,i}-P_{L,i})}{\sum_{i=1}^{n}(P_{G,i}+P_{GSP,i})} \quad (4.17)$$

Figure 4.15 illustrates the shares of controllable loads within the total daily load (in p.u.) before and after DSM for cases A.1-A.3. Due to the limited flexibility of Z and IM loads, load payback and load changes after PSO, the initial composition is not maintained in all the time steps of case A.1. Cases A.2 and A.3 result in different shares of IM loads during the day, while the shares of Z loads are very similar to case A.1 – mainly due to the similar amount of total shifted Z loads and the corresponding Z load payback in all three cases. As described in Section 4.2.2, the Z load payback happens in the three hours following a load disconnection, which may result in large shares of Z loads during the day in spite of the load disconnection during peak load times. For illustration purposes, Figure 4.16 shows demand composition before and after DSM for case A.2.

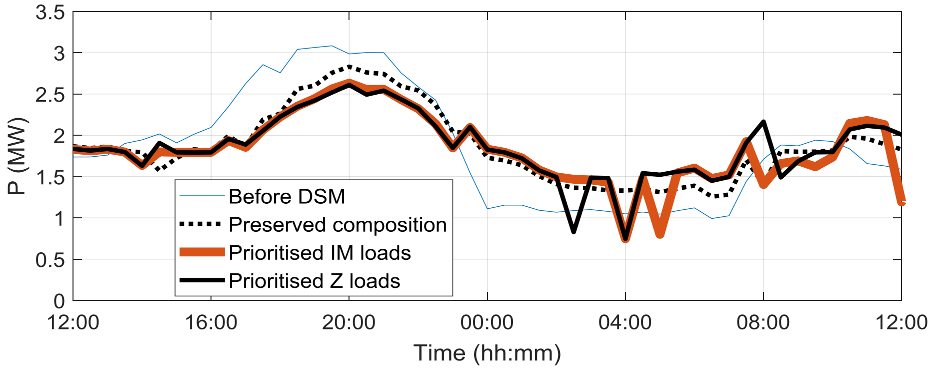


Figure 4.13 Network loading for cases A.1-A.3

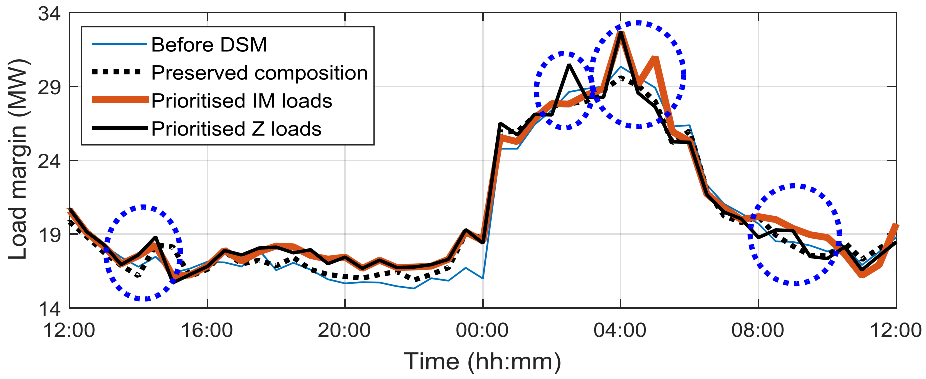


Figure 4.14 Load margin for cases A.1-A.3

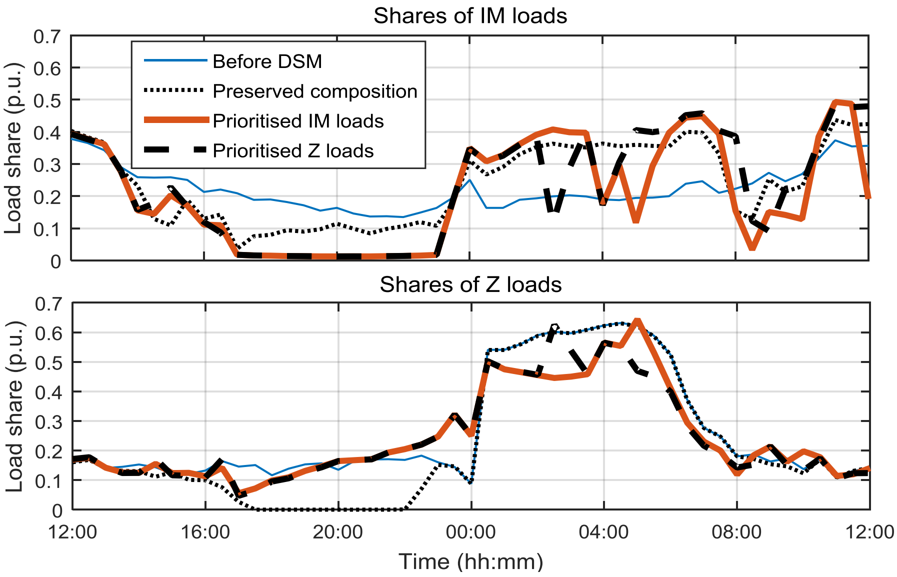


Figure 4.15 Shares of controllable demand for cases A.1-A.3

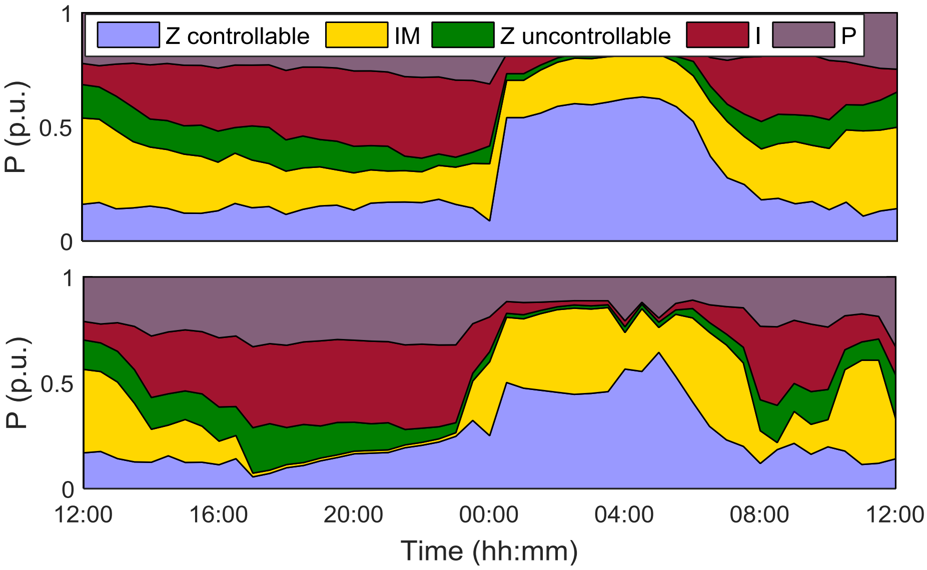


Figure 4.16 Demand composition before and after DSM for case A.2

4.4.1.3 B.1 Case with limited acceptability

Due to the limited acceptability of DSM by the end users, and consequently reduced load controllability, the loading curve after DSM could not be as flattened as in the previous three cases (Figure 4.17). Due to the reduced controllability, flexibility limit (violet line in the figure) is higher than in previous cases (Figure 4.10). As seen in Figure 4.18, both peak reduction and valley filling were limited due to limited acceptability of the end-users. Since the load was less modified, the load margin was not deteriorated at any time step (Figure 4.19). The effect of DSM on the peak to average ratio and network losses is shown in Table 4.4.

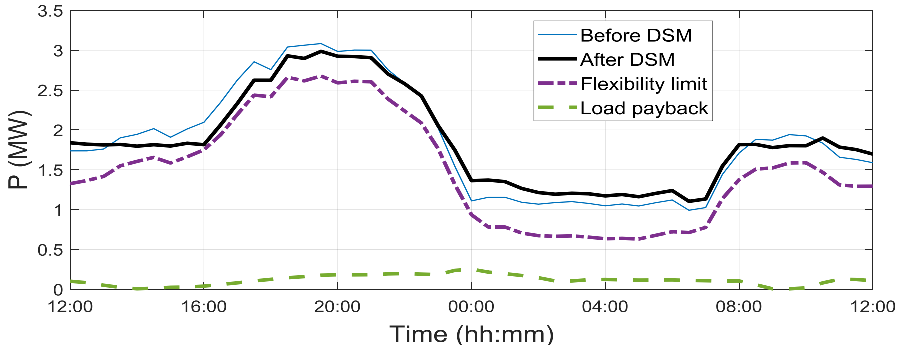


Figure 4.17 Network loading for case B.1

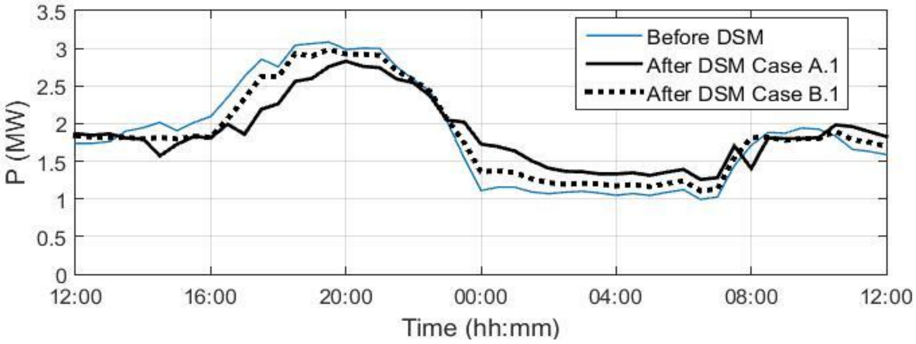


Figure 4.18 Network loading before and after DSM for cases A.1 and B.1

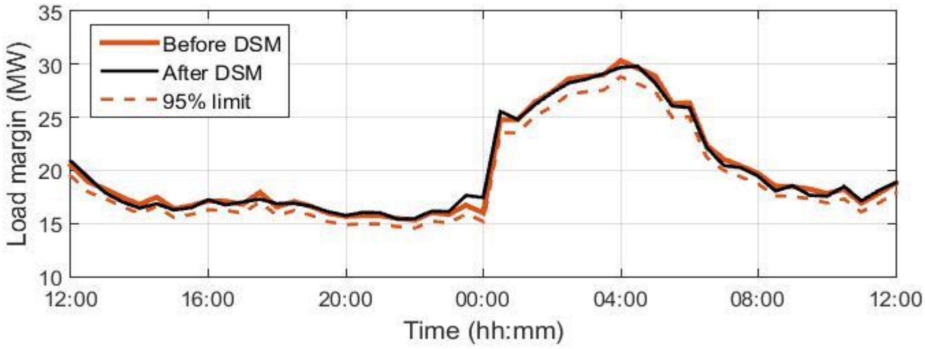


Figure 4.19 Load margin for case B1

Table 4.4 Effectiveness of DSM in case B.1

Subcase	Peak to average ratio (PAR) before/after DSM	Average daily losses/ Daily percentage losses before DSM	Average daily losses/ Daily percentage losses after DSM
B.1	1.66/ 1.60	0.046 MW/ 2.4%	0.046 MW/ 2.4%

4.4.1.4 C.1 and D.1 Overloaded system and critically loaded system

The DSM effectiveness in these two cases is shown in Table 4.5. It can be seen that the load shaping was not as successfully performed as before, due to the increased load payback, especially after the PSO algorithm, when larger amounts of load were reduced to preserve the load margin (Figure 4.20 and Figure 4.22).

Due to the increased loading in these two cases, the load margin is violated in more time steps following a DSM action, especially in the second half of the day (see Figure 4.21 and Figure 4.23), which is why PSO algorithm is run more often. It can also be seen that the minimum load margin before DSM is around 12 MW in overloaded case (Figure 4.21), and 5 MW in critically loaded system (Figure 4.23). In the base case minimum load margin is around 16 MW (Figure 4.11), i.e., 30% higher than in the overloaded case, and more than 3 times higher than in the critically loaded case.

Table 4.5 Effectiveness of DSM in cases C.1 and D.1

Subcase	Peak to average ratio (PAR) before/after DSM	Average daily losses/ Daily percentage losses before DSM	Average daily losses/ Daily percentage losses after DSM
C.1	1.66/ 1.59	0.186 MW/ 5.0%	0.159 MW/ 4.5%
D.1	1.66/ 1.65	0.790 MW/ 10.4%	0.626 MW/ 9.4%

Although the network losses after the DSM were still high in cases C.1 and D.1 (as seen in Table 4.5), the DSM program reduced the network losses significantly. Apart from the values given in Table 4.5, cumulative daily losses (given as energy in MWh) were analysed: these losses were reduced from 4.56 MWh to 3.9 MWh (14 % reduction) in the overloaded system. In the case of critically loaded system, the reduction of cumulative daily losses after DSM was even more pronounced - from 19.34 MWh to 15.35 MWh (21% reduction).

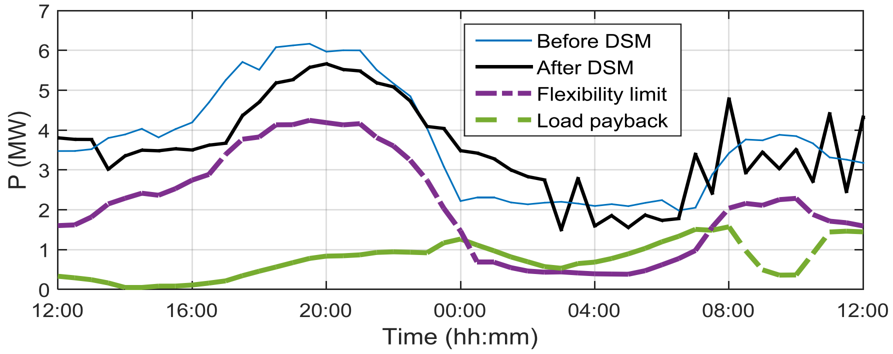


Figure 4.20 Network loading for case C.1

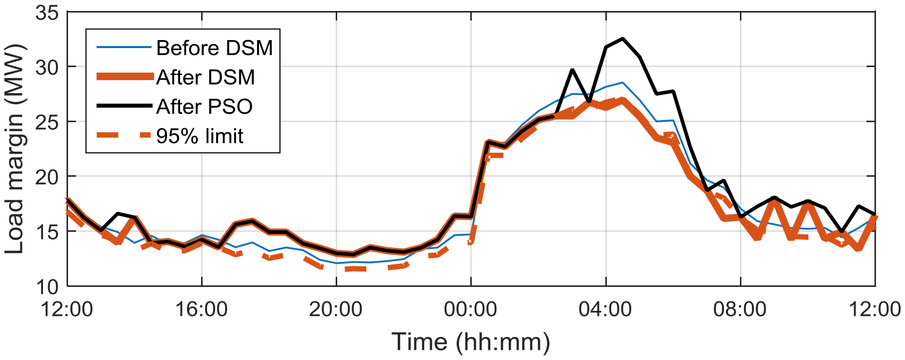


Figure 4.21 Load margin for case C.1

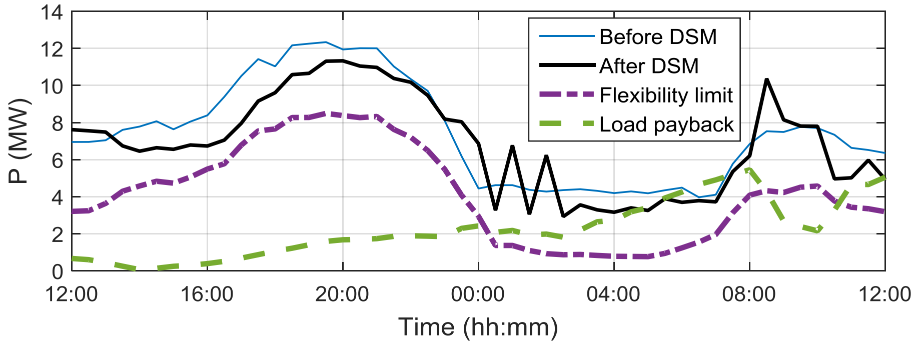


Figure 4.22 Network loading for case D.1

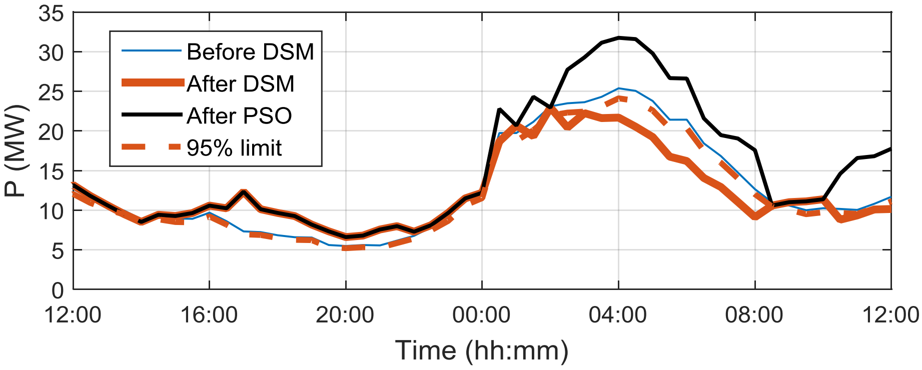


Figure 4.23 Load margin for case D.1

4.4.1.5 E.1 Neglected DSM constraints

Figure 4.24 presents the resulting load curve if DSM is performed with constant power load model instead of the composite load model (constant power model is commonly used in voltage stability analysis), and with neglected load payback and load margin. Since there is no load payback, the loading curve is not increased during valley periods. During peak (afternoon) hours all the load flexibility is harnessed, as seen in overlapped loading curve (thick black line) after DSM and flexibility limit curve (dashed violet line). Even though the peak load is successfully reduced, the load margin after DSM is deteriorated in some time steps, as seen in Figure 4.25, marked with circles. It should be noted that in a network larger than this one, the effect of neglecting intrinsic limitations of demand shift will be even more significant.

Finally, Figure 4.26 illustrates the resulting loading curve in cases A.1 and E.1, showing more successful curve flattening in case E.1. Figure 4.27 shows how “optimistic” the load reduction is when the aforementioned constraints are neglected compared to the base case (A.1), which is accounting for these constraints (namely composite load model, load payback and load margin). In both cases the changes are calculated using the same base load before DSM. During the peak load period (between 2 pm and midnight), the load reduction in most steps of case A.1 ranges between 5 and 15%, while in case E.1 the reduction is almost constantly above 30%. This shows that the expectations in load reduction may be two times higher when realistic constraints are not taken into account. Similarly, during the valley period (between midnight and 8 am), where case E.1 neglects the load payback (thus there is no change in demand), the load increase in case A.1 due to load payback reaches almost 60%. Even though this happens during the valley period in the given example, the demand increase is significant and shows the importance of load payback modelling in DSM studies. The effectiveness of DSM with respect to PAR and network losses is given in Table 4.6.

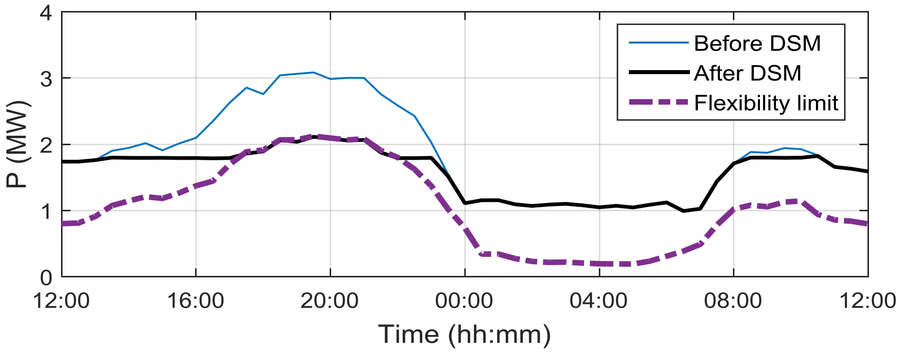


Figure 4.24 Network loading in case E.1

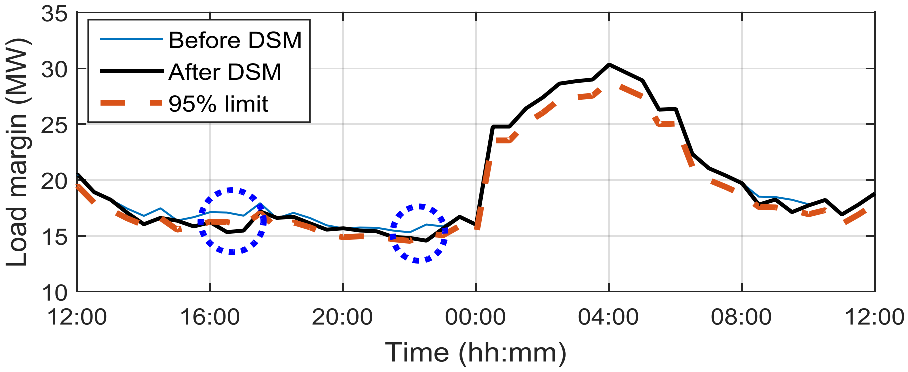


Figure 4.25 Load margin in case E.1

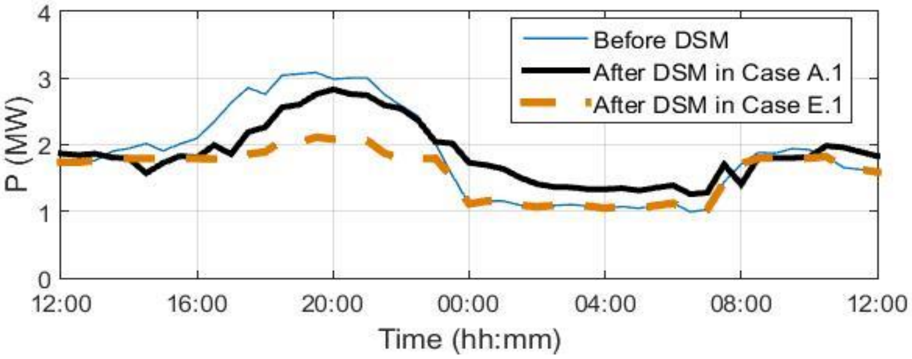


Figure 4.26 Network loading in cases A.1 and E.1

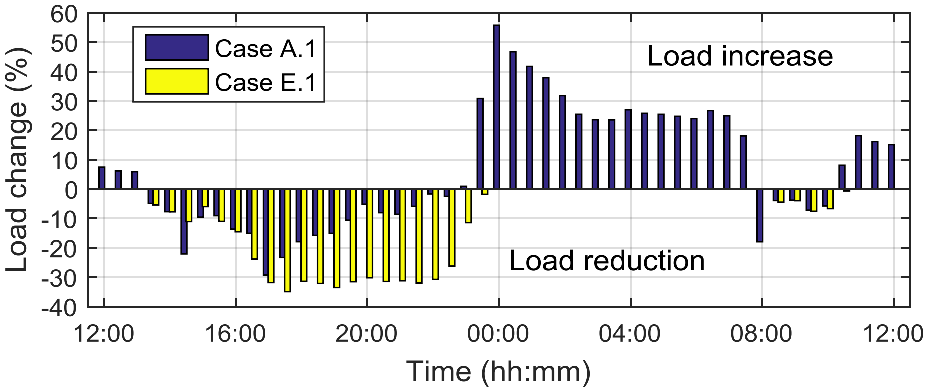


Figure 4.27 Load change after DSM in cases A.1 and E.1

Table 4.6 Effectiveness of DSM in case E.1

Subcase	Peak to average ratio (PAR) before/after DSM	Average daily losses/ Daily percentage losses before DSM	Average daily losses/ Daily percentage losses after DSM
E.1	1.66/ 1.33	0.053 MW/ 2.8%	0.041 MW/ 2.5%

4.4.1.6 Discussion

Figure 4.28 illustrates the absolute change in load margin before and after DSM for cases A.1, C.1 and D.1, where the positive values imply improvement and negative ones deterioration of the load margin. The load margin improvement (increase) is seen in most time steps, especially for overloaded and critically loaded system. Relative change in load margin, shown in Figure 4.29, reveals that with higher loading of the network (case D.1 in particular), the contribution of the DSM to load margin improvement is higher. While the relative improvement in cases A.1 and C.1 is up to 20% (compared to the load margin before DSM program), DSM in case D.1 results in improvement of up to 70% (during peak load hours).

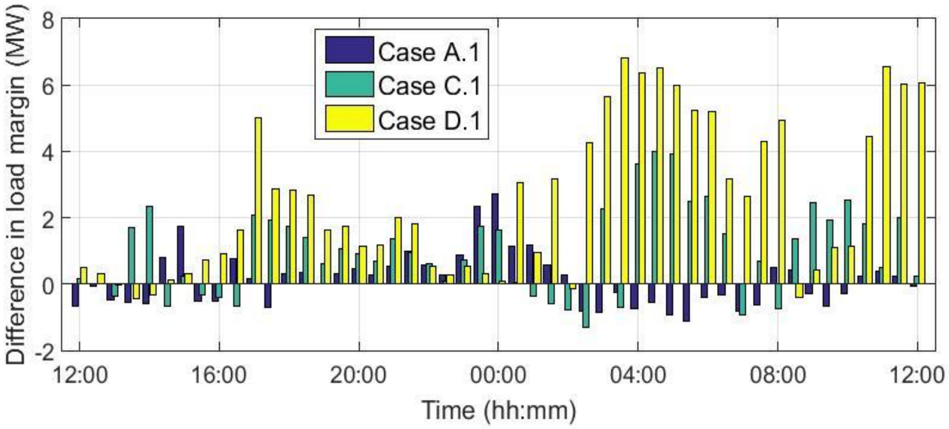


Figure 4.28 Difference in load margin (after DSM+PSO)

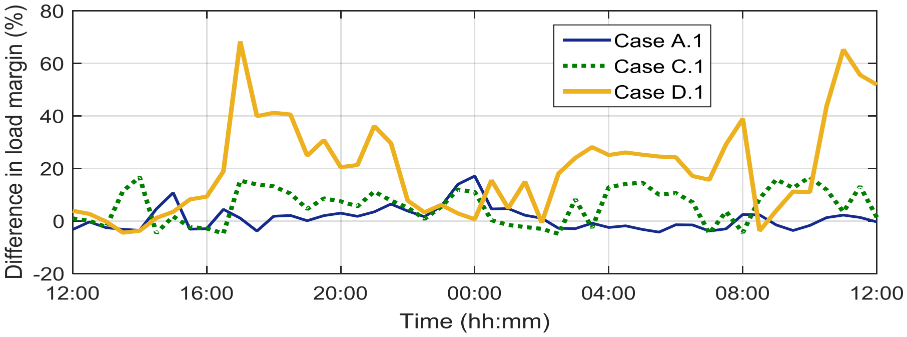


Figure 4.29 Relative difference in load margin

The results have shown that realistic constraints imposed by limited load flexibility (including limited willingness of the end-users to participate in DSM), load payback, and preservation of demand composition and loadability of the network affect the extent to which load curve can be modulated. As seen in Table 4.7 summarising all aforementioned cases, the smallest PAR, i.e., the most successful flattening of loading curve is possible only when these constraints and the composite load model are neglected (case E.1).

Reduction of network losses in the studies with base case load is also highest in case E.1, while among cases A.1-A.3 they are only slightly reduced on a daily basis. When the system is overloaded, as illustrated by cases C.1 and D.1, load shaping is limited due to larger load payback following demand shift, though reduction of losses is more notable in these case studies. This is confirmed by the results presented in Table 4.8, where relative improvement of PAR and network losses for the system with base load (case A.1) and overloaded system (cases C.1 and D.1) are shown. As already mentioned, PAR improvement is higher with base case load, while the network losses reduction is more significant in overloaded systems, reaching around 21% reduction of average daily losses in case D.1.

Table 4.7 Effectiveness of DSM in all cases

Subcase	Peak to average ratio (PAR) before/after DSM	Average daily losses/ Daily percentage losses before DSM	Average daily losses/ Daily percentage losses after DSM
A.1	1.66/ 1.51	0.046 MW/ 2.4%	0.043 MW/ 2.3%
A.2	1.66/ 1.44		0.039 MW/ 2.2%
A.3	1.66/ 1.40		0.040 MW/ 2.2%
B.1	1.66/ 1.60		0.046 MW/ 2.4%
C.1	1.66/ 1.59	0.186 MW/ 5.0%	0.159 MW/ 4.5%
D.1	1.66/ 1.65	0.790 MW/ 10.4%	0.626 MW/ 9.4%
E.1	1.66/ 1.33	0.053 MW/ 2.8%	0.041 MW/ 2.5%

Table 4.8 Relative improvement of peak to average ratio and network losses in cases A.1, C.1 and D.1

Subcase	Reduction of peak to average ratio (%)	Reduction of average daily losses (%)	Reduction of daily percentage losses (%)
A.1	**9.0**	6.5	4.2
C.1	4.2	14.5	**10.0**
D.1	0.6	**20.8**	9.6

4.4.2 Operating scenario 2: Load follow generation for isolated distribution network or microgrid

This scenario aims at illustrating possible DSM effects when different types of distribution load (namely controllable Z and IM loads) are scheduled with the objective to follow available local renewable generation. In this case demand composition is not preserved in order to observe the influence of shifting either Z or IM loads. The same distribution network model as in scenario 1 is used, with the only difference that the equivalent DG represents a solar power plant. The DG output in this case is the output of a solar plant reported in [218]. It is assumed that accurate information about available renewable generation is known day-ahead, as well as the load forecast and its composition. This allows load scheduling over a 24-hour planning horizon. The scheduling is obtained by running an OPF in Matpower [212], as defined in Section 4.3.1, with the main objective to follow the available renewable generation curve.

Two case studies are examined in this scenario. The first one observes base case loading of the network with peak daily loading of about 3 MW. In the second case study the load is scaled by factor 2 in order to illustrate overloaded operating conditions when the distribution system has a high penetration of electric vehicles (EVs), for example. It was reported in [14] that the penetration of EVs may double the distribution load, primarily at peak load hours. Load is therefore scaled only during the peak hours (between 3 pm and 12 am) in this case.

In the first case study three DSM approaches are taken (Table 4.9): *i)* IM loads are prioritised when shifting demand; *ii)* Controllable Z demand is prioritised; *iii)* Controllable Z demand is prioritised, but the payback load is scheduled in the same way as for IM loads (i.e., by scheduling load reconnection during the valley periods, as described in Section 4.2.2). The second case study only observes the DSM program with prioritisation of Z loads, to illustrate how future loading conditions may affect network performance. The main parameters analysed in the studies are the total network loading, loadability margin, network losses, and the slack bus flow before and after the DSM program. Although the methodology observes 24 hour planning horizon, some results are shown over a 36-hour period in order to demonstrate longer-term effects.

Table 4.9 Case studies

Case Study	Subcase
Base case load	(i) Prioritise IM in DSM
	(ii) Prioritise Z in DSM
	(iii) Prioritise Z in DSM with different payback load policy
Overloaded system	Prioritise Z in DSM

4.4.2.1 Case 1: Base case load

i) and ii) Prioritization of Z or IM loads

Figure 4.30 represents network loading before and after DSM, as well as the DG output. The results obtained from the two approaches are almost the same. As seen from the figure, the loading curve follows the available generation, but its ability to do so is constrained by both, limited controllability and the payback load. An interesting observation is that during the initial load shifting period, i.e., between midnight and 4am of the first day, the load margin is lower than it was before DSM (Figure 4.31), while between 5am and 10am of the first day, when the load is reconnected, the load margin is increased. In other times of load shift and during the same period of the following day the load margin is typically slightly improved (increased). This "anomaly" during the initial 2-3 hours (observable in other plots as well) can be explained by the fact that there was no "history of DSM actions" at the outset of simulations, i.e., no payback load during the first several hours.

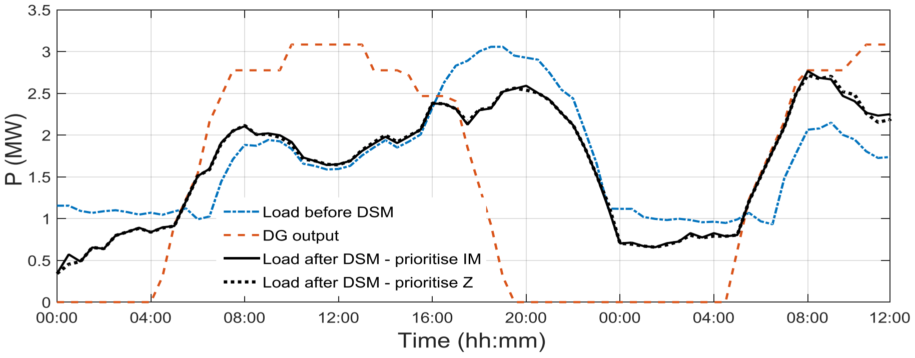


Figure 4.30 Network loading for case 1 (i) and (ii)

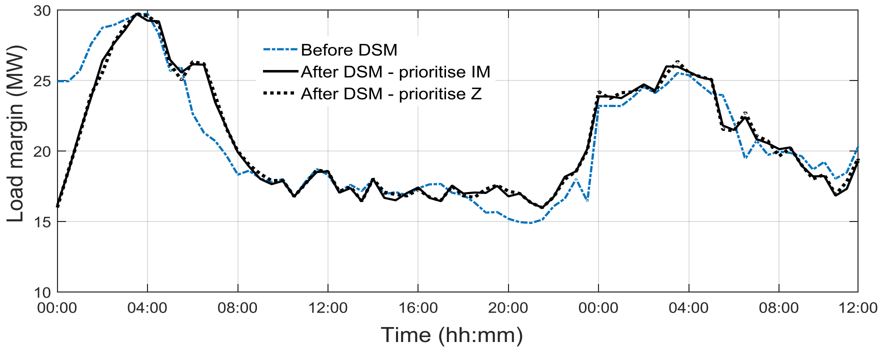


Figure 4.31 Load margin for case 1 (i) and (ii)

Losses (shown in Figure 4.32) are decreased across the whole DSM period, except between 5 and 10 am when a portion of load is being reconnected and the total load increased. Finally, the dependence of the distribution network on the upstream network is mainly reduced (reduced power import), particularly during load curtailment hours, as seen in Figure 4.33.

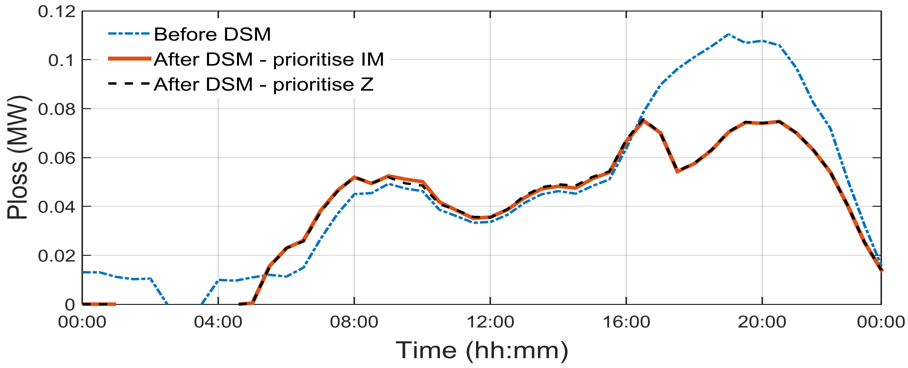


Figure 4.32 Network losses for case 1 (i) and (ii)

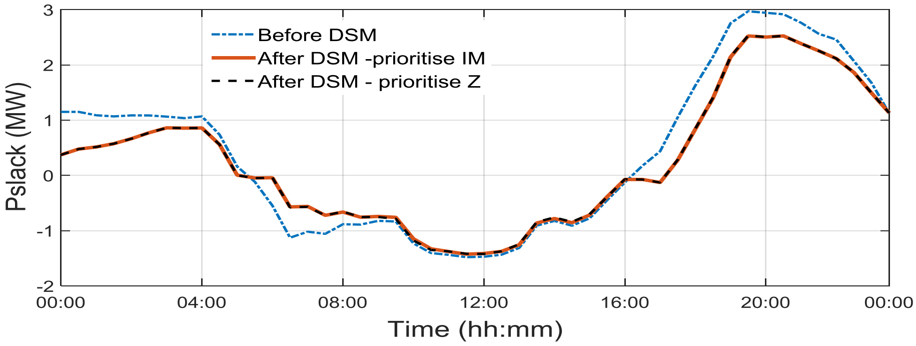


Figure 4.33 Slack bus flow for case 1 (i) and (ii)

Finally, Figure 4.34 illustrates the change in demand composition (shares of different load components in per unit of the total load in each time step) due to DSM. Clear reduction of controllable IM load component can be seen during the peak load hours, while controllable Z load does not change much due to the load payback.

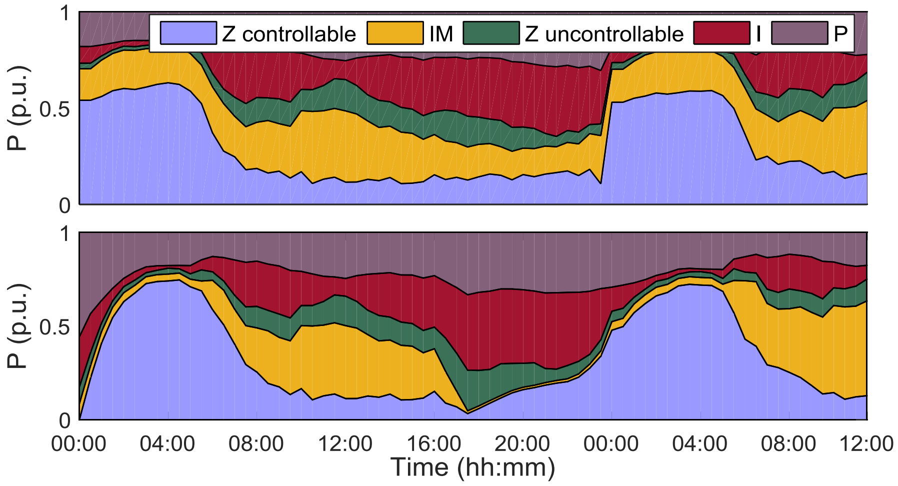


Figure 4.34 Demand composition before DSM (top) and after (bottom)

iii) Prioritization of Z loads with rescheduled payback

As seen in the base case (Figure 4.30-Figure 4.33), the results with both DSM approaches are fairly similar. The reason for this is the payback scheduling program for the Z loads, as described in Section 4.2.2. Figure 4.35 illustrates how the Z load payback follows the shift (disconnection) of Z loads in the base (original) case. It can be seen that the payback load only "shifts" the disconnected load, which means that the overall Z demand is not visibly changed, only slightly shifted in time. On the other hand, if the payback is rescheduled in the same way as the IM load (the process is described in Section 4.2.2), the Z load decrease is more prominent and better distributed during the day. This approach effectively fills the valleys of the loading curve and enables more load curtailment during the hours when it is required (e.g., between 8pm and 4am of the following day), as seen in Figure 4.36. The variation in load margin, however, is more pronounced in this case compared to base load, in particular during the reduced loading period (Figure 4.37), as the reduction in Z loads, which are beneficial for loadability, is not followed by load payback. Finally, both the reduction of losses during the period between 4pm and 12am (Figure 4.38), and the slack bus flow (i.e., dependence on the upstream network shown in Figure 4.39) are reduced with more success than using the original payback scheduling program for the Z loads.

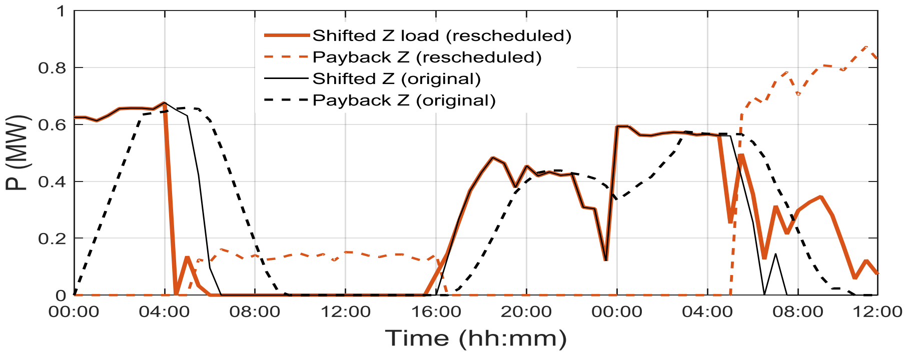


Figure 4.35 Shifted Z load and load payback

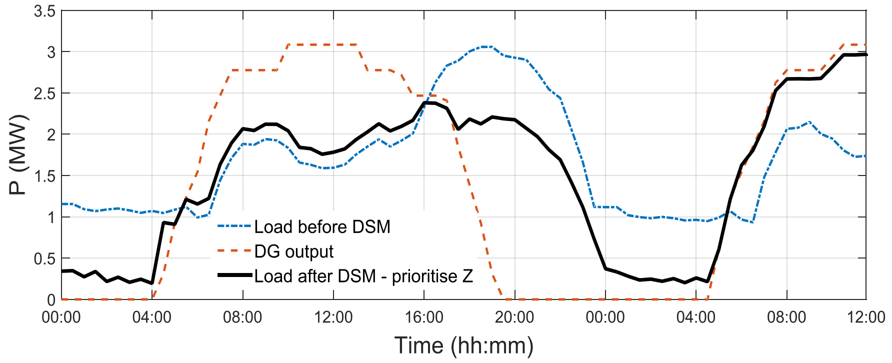


Figure 4.36 Network loading for case 1 (iii)

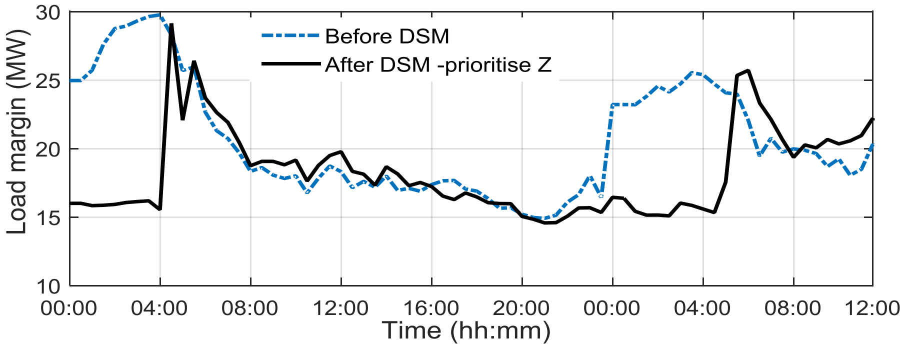


Figure 4.37 Load margin for case 1 (iii)

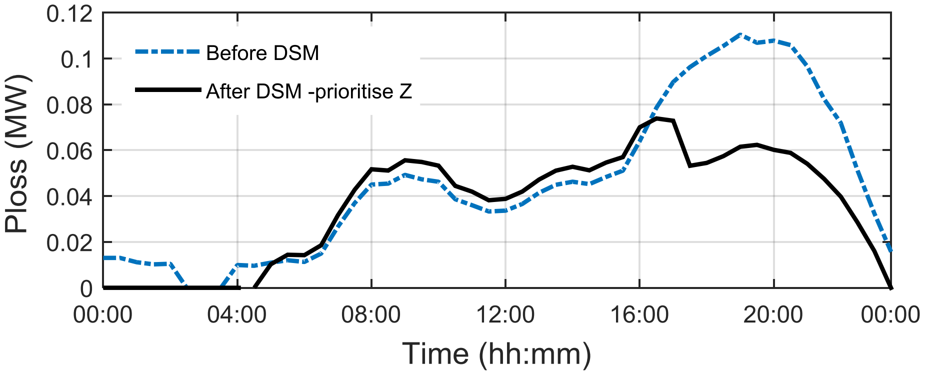


Figure 4.38 Network losses for case 1 (iii)

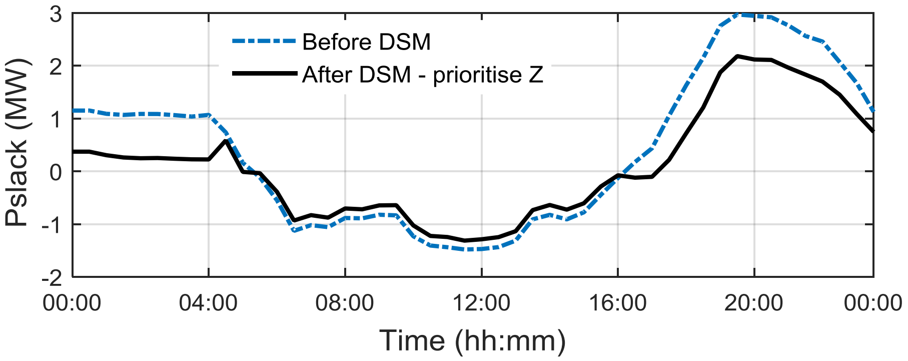


Figure 4.39 Slack bus flow for case 1 (iii)

4.4.2.2 Case 2: Overloaded system

In the case study with overloaded system, i.e., demand scaled by 2 during peak hours, the available DG cannot cover the substantial load increase during the peak hours, as seen in Figure 4.40, hence the balance needs to be provided by the external system. The load margin (see Figure 4.41) is notably improved (except during 8 am-12 pm, when the load was increased due to the payback load) in this case following the DSM, and in particular during the peak load hours when the improvement is significantly higher than in the case of base load (Figure 4.31). This improvement was expected as the originally more loaded distribution network, importing additional power from the external grid, is effectively de-loaded by the DSM action. The DSM in this case has also more prominent effect on the system losses (Figure 4.42) as they are reduced more in this than in the previous case following the DSM. As far as the import from the external grid is concerned, it can be seen (Figure 4.43) that the slack bus flow is also reduced more compared to the base case load, which further validates the effectiveness of the DSM program.

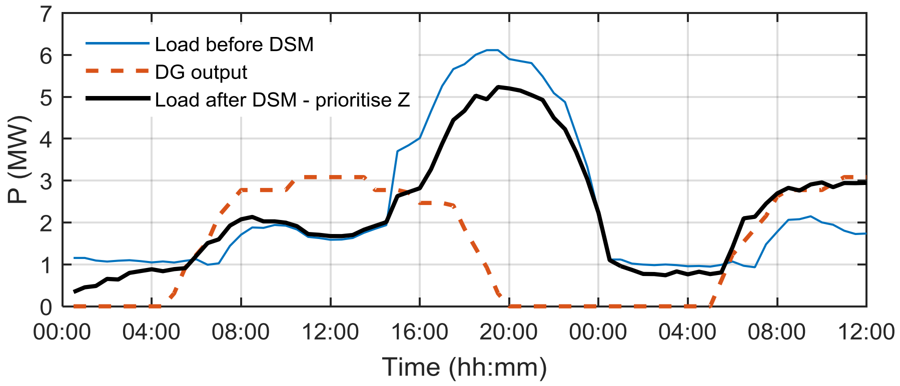


Figure 4.40 Network loading for case 2

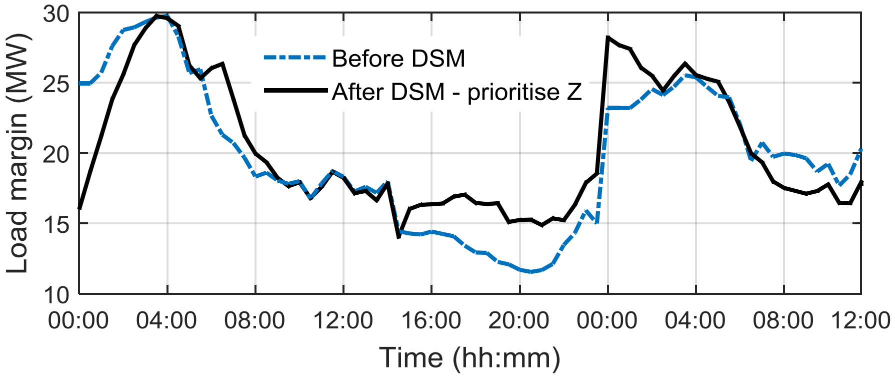


Figure 4.41 Load margin for case 2

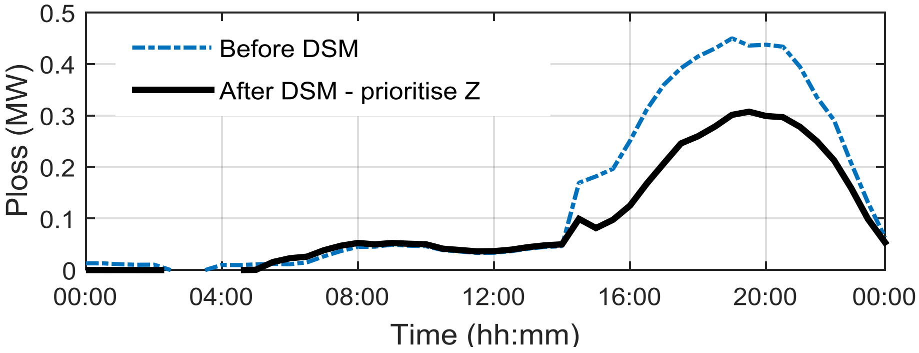


Figure 4.42 Network losses for case 2

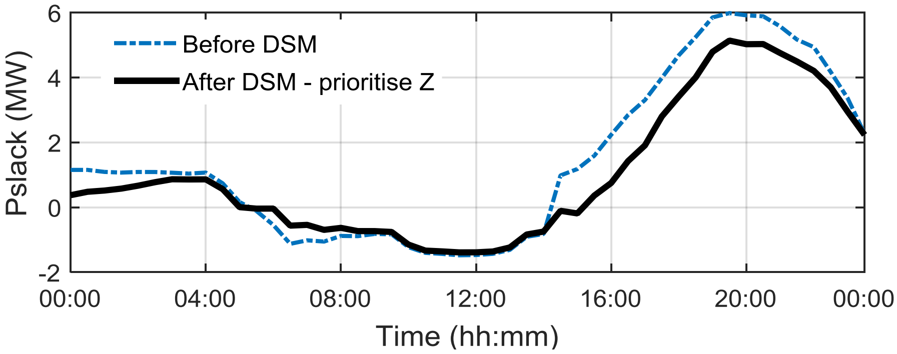


Figure 4.43 Slack bus flow for case 2

4.4.2.3 Discussion

The importance and the influence of the load reconnection on network performance following the DSM have been clearly illustrated in the previous section. Furthermore, depending on the type of load participating in DSM (in this case, ether induction motors or constant impedance loads), the load flows and voltage drops will change across the network, during both load disconnection and reconnection periods. This change in demand composition effectively, may be unexpected in some time steps of the planning horizon, which is why a detailed analysis of different DSM scenarios should be performed prior to triggering a DSM action. Information about the composition of demand, i.e., the shares of different static and dynamic load components is essential for the accuracy of these analyses. As shown on the example of prioritized Z loads with rescheduled payback, even though the target demand is met more successfully, load margin is deteriorated compared to case with linear load payback. Therefore, any wide-scale DSM action should be made only after analysing possible effects that the change in load magnitude and consequently load composition at different load buses can cause in the network.

The analysis compared cases with base case load, and those with larger load, illustrating future scenarios with high penetration of large residential loads (e.g., electric vehicles and heat pumps). The effects illustrated in this section are based on the study performed using a relatively small distribution network. The consequences and the effectiveness of DSM programs that could arise at a larger, transmission network level (including more distribution networks or large loads connected at the transmission level) could be even more important and should be carefully studied in the future for the overall power system and its stability.

4.5 Summary

This chapter presented a comprehensive methodology for optimal scheduling of distribution network loads in support of transmission network operation. The main objective of the proposed DSM program is the load profile shaping, as a balancing service to be offered to TSO while maintaining the load composition and one or more network performance indicators (the distribution network loadability in this case) to values they had prior to the DSM action, or within a pre-specified region around the original values. The influence of load modelling, limited demand flexibility (including customers' willingness to participate in the DSM program) and load payback was taken into account, illustrating the importance of considering realistic assumptions when estimating the success of a DSM program. In order to preserve the loadability of the network (in the case study illustrated in this chapter, or more generally, any other network performance indicator) after the DSM action, a trade-off between the opposing objectives must be struck and the load profile of the distribution network has to be "tailored" considering both the requirements of the network operator, and preservation of the chosen network performance indicators. The PSO–based load scheduling methodology for meeting multiple objectives of the DNO – meeting the target loading at the GSP, keeping/improving loadability of the distribution network, and maintaining the composition of demand at the GSP is the sixth original contribution of this thesis. Unlike previous work, the proposed methodology aims to schedule, optimally and simultaneously, two distinct controllable load types, namely constant impedance load (e.g., space and water heating) and induction motors (e.g., washing machines, refrigeration, HVAC), so that the load margin after the DSM program is at least maintained, if not improved. This is the seventh contribution of the thesis.

5 Conclusions and Further Work

5.1 Major conclusions

This thesis has presented the results of the research performed in two main areas: decomposition of aggregated demand for demand-side flexibility assessment, and optimised DSM. The main aim of the research was to develop a methodology for multi-objective DSM in distribution network in support of transmission network operation, relying on the existence of a certain number of SMs with sub-metering technologies and application of data analytics methods, namely ANN.

The summary of chapters and the main findings of the research within are given in the following sections.

5.1.1 Chapter 1 Introduction

This chapter introduced the main research areas presented in this thesis, with a critical overview of the past work in these areas. The need for enhanced demand observability was emphasised as one of the main enablers for reliable DR programs in the evolving smart grids. Different types of DR programs were observed, as well as different approaches for the assessment of demand-side flexibility. Finally, potential for transmission network-level services provided by wide-scale DR was analysed, from both research and practical, industrial perspectives. Following the overview of past work, main aims and objectives of the research were defined, together with the main contributions of the thesis.

5.1.2 Chapter 2 The Need for and Application of Data Analytics in Distribution System Studies

This chapter provided an overview of the data mining methods commonly used in distribution network studies. Special attention was given to introducing possibilities of text mining in power system studies, as this area has been mostly unexplored. In addition, the chapter investigated present and future data requirements for enhanced operation and control of distribution networks, as well as the extent to which smart meters, whose proliferation is constantly growing, could meet these requirements. Both the benefits and challenges related to smart meter technologies were presented, following a detailed analysis of smart meter specifications and smart meter rollout in different countries in Europe. Finally, a case study illustrating application of simple data mining methods to a real distribution utility database was presented. The results showed that historical data can be used to model predictive tools that can be used as decision support in asset management. The main value of data mining in this example is that it enables prediction of failures and planning of asset maintenance/replacement based only on historical data and without additional cost for the utility.

5.1.2.1 Main findings

Data requirements and data analytics in power network studies

With the proliferation of ICT technologies in distribution networks, the need for data analytics methods has been raised. The two main values of data analytics methods in power system analysis are extraction of useful knowledge from big data and forecasting based on historical observations. *The overview and critical comparison of different data analytics methods, including text mining, for application in distribution system studies presents the first original contribution of this thesis. The identification of data needs in present and future distribution networks, and the extent to which smart metering can facilitate collection of these data, represents the second original contribution of this thesis.*

Application of data mining to a distribution utility database

As an example of forecasting based on historical observations using data mining, the case study given in the chapter showed that candidate assets for preventive or corrective maintenance can be identified based only on their physical characteristics and using regression models built on historical data, without investing in additional monitoring systems. Furthermore, as an example of prioritising data using data mining, regression tree, which was used in the analysis, showed that the number of faults at the substation level was not highly correlated with the total number of customers supplied by the substation. This infers that this type of data does not have to be collected for confident assessment of the expected number of faults at a substation based on feeder characteristics.

5.1.3 Chapter 3 Advanced Demand Profiling

In this chapter a methodology was developed for aggregated demand decomposition based on limited SM data with sub-metering technologies and the application of ANN. The resulting demand composition provides information about the estimated/forecast shares of different load categories and of controllable and non-controllable load within the total aggregated demand. In addition, realistic challenges related to smart meter data streams were accounted for, namely missing data and data arriving at different time steps. Following this, two data pre-processing methods were compared, linear interpolation and weight adjusted k-nearest neighbour method. Furthermore, as a solution to commonly missing reactive demand data, a probabilistic approach was developed for deriving reactive demand measurements based on active demand and probabilistic range of power factor values. Although the focus in this chapter was on day-ahead forecasting, the proposed methodology can be equally used for either real-time (estimation) or forecasting applications. The methodology was first tested on realistic statistics-based dataset, and further validated on a dataset from a real pilot site. The ANN-based approach was also compared to a time-series method, i.e., ARIMA, showing better performance.

5.1.3.1 Main findings

Advanced demand decomposition

With the suggested approach for demand decomposition, even with missing data and partial coverage of the end-users with SMs having sub-metering enabled, the confidence of demand decomposition is high. Therefore, only a limited number of SMs with sub-metering in an area would suffice for a confident estimation/prediction of aggregated demand composition. As the methodology relies on aggregated and relatively small historical data (from the most recent week), the proposed approach does not require significant data storing and communication resources, which brings additional savings to the network operator. A comparative analysis also showed that additional types of training (input) data, such as weather and type of the day, do not improve the accuracy of the algorithm. *The methodology for aggregated demand decomposition using limited number of smart meters with enabled sub-metering represents the third original contribution of the thesis.*

Probabilistic derivation of reactive load data

As a solution to the lack of reactive load measurements at the end-users' point (which is a common deficiency of some types of SMs), a method for obtaining probabilistic aggregated reactive load data was developed. Reactive load data is necessary for both appropriate modelling of demand and power factor at each load bus, and as an input to the ANN-based demand decomposition algorithm. The approach requires real power data, both total and decomposed to load category level, and derives reactive demand data probabilistically, based on probabilistic modelling of PF of individual load categories. *This methodology represents the fourth original contribution of the thesis.*

Graphical user interface for advanced demand profiling

A graphical user interface (GUI) was developed for representing aggregated demand forecast and demand decomposition, as a support tool for DSM planning. The GUI can be used in a control room by the network operator who will be able to forecast total demand and its flexibility during the planning horizon (in this case, 24 hours). *Development of the graphical user interface represents the fifth original contribution of this thesis.*

5.1.4 Chapter 4 Multi-objective Demand Side Management at Distribution Network Level

A methodology for optimal scheduling of distribution network loads in support of transmission network operation was developed in this chapter. The aim of the suggested DSM program is to meet the predefined loading curve, maintaining at the same time demand composition and distribution network loadability, i.e., the load margin, as the chosen network performance indicator. The importance of realistic assumptions related to demand, namely the appropriate load model, limited demand flexibility, modelling of the load payback and end-users willingness to participate in DSM, is illustrated by analysing the effect these factors may have on the resulting loading curve after DSM. A particle swarm optimisation based algorithm for demand scheduling was developed in order to meet multiple objectives of the network operator – meeting the target loading at the GSP, keeping/improving loadability of the distribution network, and maintaining the composition of demand at the GSP to preserve the loadability of the network after the DSM action.

5.1.4.1 Main findings

Multi-objective DSM

The optimised DSM program proposed in this thesis has as the main objective load profile modulation, as a balancing service offered by the DSO to the TSO, while at the same time maintaining the load composition at GSP and one or more network performance indicators (the distribution network loadability in this case) to values they had prior to the DSM action. Realistic aspects of demand, namely static and dynamic components of the load, limited demand flexibility (including customers’ willingness to participate in the DSM program) and load payback, were taken into account. Results of the analyses shown in several case studies have proven that meeting the predefined target load curve is limited by the aforementioned constraints, including intrinsic nature of demand and its flexibility on one hand, and the preservation of network performance on the other. Therefore, the DSM program has to be “tailored” considering both the requirements of the network operator, and preservation of the chosen network performance indicators. *The development of the proposed multi-objective DSM program relying on particle swarm optimisation is the sixth original contribution of this thesis.*

Optimal scheduling of load types

Instead of scheduling constant power load, as in most cases in previous work on DSM, the proposed DSM methodology schedules in an optimal way two controllable load types, namely constant impedance loads and induction motor loads. These two load types show different static and dynamic behaviour, which may influence both steady state and dynamic performance of the power network. Therefore, the optimal scheduling is performed to maintain or improve static voltage stability indicator (network loadability) and prevent undesirable dynamic response of demand in case of a disturbance. *The optimal scheduling of more than one controllable load as a part of the overall DSM program represents the seventh original contribution of the thesis.*

5.2 Further work

Some of the research problems identified during the course of this research could not be addressed and presented in this thesis, mainly due to the limited time. These research areas, detailed below, will be considered in future work.

5.2.1 The use of smart meter data

Apart from the analysis of benefits for distribution network operators coming from SM data, further research should investigate other possible applications of SM data mining, especially when combined with other types of data, such as weather, sociodemographic data, transport, etc. Weather data and advanced data mining methods can complement SM data in individual household load forecasting. Furthermore, correlation between SM data and sociodemographic data can provide more significant insight into end-users' behaviour, which can be highly beneficial for designing different DR programs tailored based on the knowledge about this behaviour. Historical data about the demand consumption provided by SMs and data about social events in the area can be used to predict high demand in certain areas of the distribution network, for example. Similarly, smart metering of EV charging can be used to foresee possible traffic congestion problems.

In addition, there is still an untapped potential for application of text mining for improved understanding of end-users behaviour and daily load profiling. For example, text mining of on-line reports or information from social media can give an insight into changes in customer preferences with respect to types of load, usage of EVs, installation of home energy storage systems, etc., which can be valuable for understanding the changes in daily load profiles recorded by SMs.

5.2.2 Demand decomposition

Methodology for demand decomposition should include other load types, such as EVs and heat pumps, as the share of these loads is constantly growing in distribution network. Furthermore, the ANN training process should be extended to allow for inclusion of historical data with DR events. It is important to distinguish between intrinsic changes in demand due to natural behaviour of the end-users, and the changes resulting from DR programs. Finally, the impact of distributed generation on results of demand decomposition should also be accounted for, as it effectively changes the loading curve, if it is partly supplied by distributed generation.

The application of other types of data mining, for example deep learning neural networks or recurrent ANN, for demand decomposition should be investigated. The aim is to improve generalisation of the demand decomposition tool, and enable more confident results in cases when the training and testing data are not necessarily from the same consumption area.

5.2.3 Modelling of DSM

Modelling the distribution network as an unbalanced system should be considered in multi-objective DSM, primarily due to their intrinsic unbalance and the fact that larger loads, such as EVs and heat pumps, may introduce larger unbalances. In addition, the optimisation problem should include control of storage and EV charging.

Furthermore, multi-period optimisation should be investigated instead of sequential optimisation presented in this thesis, as it introduces time as one of the variables, and allows for optimal scheduling of demand taking into account both previous and future time steps of the planning horizon. This approach could enable more efficient and controlled shaping of the load payback, preventing overloading due to reconnection of shifted demand.

Load payback was modelled in a simplified way in this thesis, mainly to illustrate possible network operation and DSM issues arising from demand shifting of different load components. More realistic modelling should be performed in the future, as well as smarter, i.e., more optimal ways of scheduling (controlling) load reconnection. Special attention should be given to thermostatically controlled loads, whose operation is highly affected by temperature.

Further work should also consider better modelling of customers' willingness to participate in DR, both in terms of number and geographical location, as well as the impact of this on the effectiveness of a DSM program.

The proposed multi-objective DSM approach should be further extended and validated on a system comprising transmission network and one or more connected large distribution networks or large industrial users. This would enable a more realistic assessment of the contribution of DSM to the transmission system operation.

Finally, to fully appreciate the potential of DSM, other network performance indicators (individually, or as a combination of a few), apart from loadability, at transmission or distribution level, should be incorporated into the optimised load scheduling methodology. These may include network losses or different aspects of system stability. With more performance indicators included in DSM, more benefits could be derived for the network, as well as for the individual customers participating in DSM program. Considering that the computational complexity of the task would rapidly increase with more network buses and more parameters and performance indicators to consider, other dedicated and potentially more efficient optimisation approaches for load scheduling should be investigated.

References

[1] "Electric Power System Flexibility: Challenges and Opportunities," Electric Power Research Institute, Palo Alto, California, 2016.

[2] C. Kang, Y. Wang, Y. Xue, G. Mu, and R. Liao, "Big Data Analytics in China's Electric Power Industry: Modern Information, Communication Technologies, and Millions of Smart Meters," *IEEE Power and Energy Magazine*, vol. 16, pp. 54-65, 2018.

[3] F. Li, R. Li, Z. Zhang, M. Dale, D. Tolley, and P. Ahokangas, "Big Data Analytics for Flexible Energy Sharing: Accelerating a Low-Carbon Future," *IEEE Power and Energy Magazine*, vol. 16, pp. 35-42, 2018.

[4] K. Samarakoon, J. Ekanayake, and N. Jenkins, "Reporting Available Demand Response,", *IEEE Transactions on Smart Grid,* vol. 4, pp. 1842-1851, 2013.

[5] M. Woolf, T. Ustinova, E. Ortega, H. O'Brien, P. Djapic, and G. Strbac, "Distributed generation and demand side response," *Report A7 for the "Low Carbon London" LCNF project: Imperial College London,* 2014.

[6] M. Pipattanasomporn, M. Kuzlu, S. Rahman, and Y. Teklu, "Load Profiles of Selected Major Household Appliances and Their Demand Response Opportunities," *IEEE Transactions on Smart Grid,* vol. 5, pp. 742-750, 2014.

[7] W. Abrahamse, S. Darby, and K. McComas, "Communication Is Key: How to Discuss Energy and Environmental Issues with Consumers," *IEEE Power and Energy Magazine*, vol. 16, pp. 29-34, 2018.

[8] "Digest of United Kingdom Energy Statistics 2015," Department of Energy and Climate Change, [Online].Available:https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/450302/DUKES_2015.pdf.

[9] B. J. Claessens, P. Vrancx, and F. Ruelens, "Convolutional Neural Networks for Automatic State-Time Feature Extraction in Reinforcement Learning Applied to Residential Load Control," *IEEE Transactions on Smart Grid*, vol. 9, pp. 3259-3269, 2018.

[10] A. Brooks, E. Lu, D. Reicher, C. Spirakis, and B. Weihl, "Demand Dispatch," *IEEE Power and Energy Magazine*, vol. 8, pp. 20-29, 2010.

[11] "CEER Status Review on European Regulatory Approaches Enabling Smart Grids Solutions ("Smart Regulation") " Council of European Energy Regulators 2014, [Online].Available: https://www.ceer.eu/documents/104400/-/-/f83fc0d2-bff9-600b-3e0f-14eccad7a8d8.

[12] A. Jalali and M. Aldeen, "Modified modal analysis approach for distribution power systems," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe),* Torino, Italy, 2017, pp. 1-6.

[13] Y. Zhu, "Ranking of Power System Loads Based on Their Influence on Power System Stability," First Year Transfer Report, The University of Manchester, 2016.

[14] "Preparing UK Electricity Networks for Electric Vehicles," [Online].Available: https://es.catapult.org.uk/wp-content/uploads/2018/07/Preparing-UK-Electricity-Networks-for-Electric-Vehicles-FINAL.pdf, 2018.

[15] "Future Energy Scenarios," National Grid, 2018, [Online].Available: http://fes.nationalgrid.com/media/1363/fes-interactive-version-final.pdf.

[16] Y. Xu, "Probabilistic Estimation and Prediction of the Dynamic Response of the Demand at Bulk Supply Points," PhD thesis, School of Electrical and Electronic Engineering, University of Manchester, 2015.

[17] Y. V. Makarov, D. J. Hill, and J. V. Milanovic, "Effect of load uncertainty on small disturbance stability margins in open-access power systems," in *System Sciences, 1997, Proceedings of the Thirtieth Hawaii International Conference on*, 1997, pp. 648-657.

[18] E. Carpaneto and G. Chicco, "Probabilistic characterisation of the aggregated residential load patterns," *Generation, Transmission & Distribution, IET,* vol. 2, pp. 373-382, 2008.

[19] C. A. Canizares, "Voltage stability assessment: concepts, practices and tools," *IEEE/PES power system stability subcommittee special publication,* 2002.

[20] R. Al Abri, E. F. El-Saadany, and Y. M. Atwa, "Optimal placement and sizing method to improve the voltage stability margin in a distribution system using distributed generation," *IEEE Transactions on Power Systems,* vol. 28, pp. 326-334, 2013.

[21] H. Hedayati, S. A. Nabaviniaki, and A. Akbarimajd, "A Method for Placement of DG Units in Distribution Networks," *IEEE Transactions on Power Delivery,* vol. 23, pp. 1620-1628, 2008.

[22] M. Ettehadi, H. Ghasemi, and S. Vaez-Zadeh, "Voltage Stability-Based DG Placement in Distribution Networks," *IEEE Transactions on Power Delivery,* vol. 28, pp. 171-178, 2013.

[23] "Microgrid Stability Definitions, Analysis, and Modeling," IEEE-PES Task Force on Microgrid Stability Analysis and Modeling,2018.

[24] Pecan Street Inc. Dataport 2017 [Online]. Available: http://www.pecanstreet.org/

[25] D. Gerbec, S. Gasperic, and F. Gubina, "Determination and allocation of typical load profiles to the eligible consumers," in *Power Tech Conference Proceedings, 2003 IEEE* Bologna, Italy, 2003, p. 5 pp. Vol.1.

[26] "Load Profiles and Their Use in Electricity Settlement," Elexon2013, [Online].Available: https://www.elexon.co.uk/wp-content/uploads/2013/11/load_profiles_v2.0_cgi.pdf.

[27] D. S. Kirschen, "Demand-side view of electricity markets," *IEEE Transactions on Power Systems,* vol. 18, pp. 520-527, 2003.

[28] A. J. Urquhart and M. Thomson, "Impacts of Demand Data Time Resolution on Estimates of Distribution System Energy Losses," *Power Systems, IEEE Transactions on,* vol. 30, pp. 1483-1491, 2015.

[29] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *Power Systems, IEEE Transactions on,* vol. 21, pp. 933-940, 2006.

[30] G.Chicco, "A Multi-faceted View on the Characterisation of Electrical Demand," *Invited talk at The University of Manchester,* 2015.

[31] V. Rigoni, L. F. Ochoa, G. Chicco, A. Navarro-Espinosa, and T. Gozel, "Representative Residential LV Feeders: A Case Study for the North West of England," *Power Systems, IEEE Transactions on,* vol. PP, pp. 1-13, 2015.

[32] I. A. Sajjad, G. Chicco, and R. Napoli, "Definitions of demand flexibility for aggregate residential loads," *IEEE Transactions on Smart Grid,* vol. 7, pp. 2633-2643, 2016.

[33] W. Kong, Y. Xu, Z. Dong, D. J. Hill, J. Ma, and C. Lu, "An extended prototypical smart meter architecture for demand side management," in *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on*, 2015, pp. 1008-1013.

[34] A. Reinhardt, P. Baumann, D. Burgstahler, M. Hollick, H. Chonov, M. Werner, and R. Steinmetz, "On the accuracy of appliance identification based on distributed load metering data," in *Sustainable Internet and ICT for Sustainability (SustainIT),* Pisa, Italy, 2012, pp. 1-9.

[35] G. S. Ledva, Z. Du, L. Balzano, and J. L. Mathieu, "Disaggregating Load by Type from Distribution System Measurements in Real Time," in *Energy Markets and Responsive Grids*, ed: Springer, 2018, pp. 413-437.

[36] D. Srinivasan, W. S. Ng, and A. C. Liew, "Neural-network-based signature recognition for harmonic source identification," *IEEE Transactions on Power Delivery,* vol. 21, pp. 398-405, 2006.

[37] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, 2011, pp. 59-62.

[38] J. Kelly and W. Knottenbelt, "Neural nilm: Deep neural networks applied to energy disaggregation," in *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, Seoul, South Korea, 2015, pp. 55-64.

[39] Y. Xu and J. V. Milanović, "Artificial-Intelligence-Based Methodology for Load Disaggregation at Bulk Supply Point," *IEEE Transactions on Power Systems,* vol. 30, pp. 795-803, 2015.

[40] J. Liang, S. K. Ng, G. Kendall, and J. W. Cheng, "Load signature study—Part II: Disaggregation framework, simulation, and applications," *IEEE Transactions on Power Delivery,* vol. 25, pp. 561-569, 2010.

[41] X. Tang, K. N. Hasan, J. V. Milanović, K. Bailey, and S. J. Stott, "Estimation and Validation of Characteristic Load Profile through Smart Grid Trials in a Medium Voltage Distribution Network," *IEEE Transactions on Power Systems,* vol. 33, pp. 1848-1859, 2018.

[42] A. J. Collin, G. Tsagarakis, A. E. Kiprakis, and S. McLaughlin, "Development of low-voltage load models for the residential load sector," *IEEE Transactions on Power Systems,* vol. 29, pp. 2180-2188, 2014.

[43] P. Aristidou, G. Valverde, and T. Van Cutsem, "Contribution of distribution network control to voltage stability: A case study," *IEEE Transactions on Smart Grid,* vol. 8, pp. 106-116, 2017.

[44] P. Kundur, N. J. Balu, and M. G. Lauby, *Power system stability and control* vol. 7: McGraw-hill New York, 1994.

[45] X. Tang and J. V. Milanović, "Assessment of the impact of demand side management on power system small signal stability," in *2017 IEEE Manchester PowerTech*, 2017, pp. 1-6.

[46] J. V. Milanović and Y. Xu, "Methodology for Estimation of Dynamic Response of Demand Using Limited Data," *IEEE Transactions on Power Systems,* vol. 30, pp. 1288-1297, 2015.

[47] A. P. A. d. Silva, C. Ferreira, A. C. Z. d. Souza, and G. Lambert-Torres, "A new constructive ANN and its application to electric load representation," *IEEE Transactions on Power Systems,* vol. 12, pp. 1569-1575, 1997.

[48] "Modelling and Aggregation of Loads in Flexible Power Networks," CIGRE WG C4.605 (566), ISBN: 978-2-85873-261-6, February 2014.

[49] J. Aghaei, M. I. Alizadeh, A. Abdollahi, and M. Barani, "Allocation of demand response resources: toward an effective contribution to power system voltage stability," *IET Generation, Transmission & Distribution,* vol. 10, pp. 4169-4177, 2016.

[50] M. H. Albadi and E. F. El-Saadany, "A summary of demand response in electricity markets," *Electric Power Systems Research*, vol. 78, pp. 1989-1996, 2008.

[51] T. Logenthiran, D. Srinivasan, and T. Z. Shun, "Demand Side Management in Smart Grid Using Heuristic Optimization," *IEEE Transactions on Smart Grid*, vol. 3, pp. 1244-1252, 2012.

[52] A. Agnetis, G. Dellino, G. De Pascale, G. Innocenti, M. Pranzo, and A. Vicino, "Optimization models for consumer flexibility aggregation in smart grids: The ADDRESS approach," in *2011 IEEE First International Workshop on Smart Grid Modeling and Simulation (SGMS)*, Brussels, Belgium, 2011, pp. 96-101.

[53] K. Christakou, "A unified control strategy for active distribution networks via demand response and distributed energy storage systems," *Sustainable Energy, Grids and Networks*, vol. 6, pp. 1-6, 2016.

[54] B. P. Hayes, "Distributed generation and demand side management: Applications to transmission system operation," PhD Thesis. University of Edinburgh, 2013.

[55] I. Cobelo, "Active Control of Distribution Networks," PhD Thesis, The University of Manchester, 2005.

[56] "Aggregators - Barriers and External Impacts," OFGEM 2016, [Online].Available: https://www.ofgem.gov.uk/publications-and-updates/aggregators-barriers-and-external-impacts-report-pa-consulting.

[57] H. Zhong, Q. Xia, C. Kang, M. Ding, J. Yao, and S. Yang, "An Efficient Decomposition Method for the Integrated Dispatch of Generation and Load," *IEEE Transactions on Power Systems*, vol. 30, pp. 2923-2933, 2015.

[58] M. Muratori and G. Rizzoni, "Residential Demand Response: Dynamic Energy Management and Time-Varying Electricity Pricing," *IEEE Transactions on Power Systems*, vol. 31, pp. 1108-1117, 2016.

[59] H. Zhong, L. Xie, and Q. Xia, "Coupon incentive-based demand response: Theory and case study," *IEEE Transactions on Power Systems*, vol. 28, pp. 1266-1276, 2013.

[60] "2011 Demand Response Availability Report," North American Electric Reliabiity Corporation,March 2013.

[61] V. Levi, "Demand Response Mechanisms and Network Support Services," *Presentation at the University of Manchester*, 2018.

[62] D. S. Callaway and I. A. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, pp. 184-199, 2011.

[63] L. Steg, R. Shwom, and T. Dietz, "What drives energy consumers?: Engaging people in a sustainable energy transition," *IEEE Power and Energy Magazine*, vol. 16, pp. 20-28, 2018.

[64] E. van der Werff, J. Thogersen, and W. B. de Bruin, "Changing Household Energy Usage: The Downsides of Incentives and How to Overcome Them," *IEEE Power and Energy Magazine*, vol. 16, pp. 42-48, 2018.

[65] A. De Paola, D. Angeli, and G. Strbac, "Convergence and optimality of a new iterative price-based scheme for distributed coordination of flexible loads in the electricity market," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Melbourne, Australia, 2017, pp. 1386-1393.

[66] F. Elghitani and W. Zhuang, "Aggregating a large number of residential appliances for demand response applications," *IEEE Transactions on Smart Grid*, vol. 9, pp. 5092-5100, 2018.

[67] P. Du and N. Lu, "Appliance Commitment for Household Load Scheduling," *IEEE Transactions on Smart Grid*, vol. 2, pp. 411-419, 2011.

[68] "The Integrated Grid - Realizing the Full Value of Central and Distributed Energy Resources," EPRI, 2014, [Online].Available: https://www.energy.gov/sites/prod/files/2015/03/f20/EPRI%20Integrated%20Grid021014.pdf

[69] Y. Xiang, L. Junyong, Y. Wei, and C. Huang, "Active energy management strategies for active distribution system," *Journal of Modern Power Systems and Clean Energy*, vol. 3, pp. 533-543, 2015.

[70] D. Wang, H. Jia, C. Wang, N. Lu, M. Fan, W. Miao, and Z. Liu, "Performance evaluation of controlling thermostatically controlled appliances as virtual generators using comfort-constrained state-queueing models," *IET Generation, Transmission & Distribution*, vol. 8, pp. 591-599, 2014.

[71] J. Medina, N. Muller, and I. Roytelman, "Demand response and distribution grid operations: Opportunities and challenges," *IEEE Transactions on Smart Grid*, vol. 1, pp. 193-198, 2010.

[72] O. Ma, K. Cheung, D. J. Olsen, N. Matson, M. D. Sohn, C. M. Rose, J. H. Dudley, S. Goli, S. Kiliccote, and P. Cappers, "Demand response and energy storage integration study," National Renewable Energy Lab.(NREL), Golden, CO (United States)2016.

[73] S. Kiliccote, D. Olsen, M. D. Sohn, and M. A. Piette, "Characterization of demand response in the commercial, industrial, and residential sectors in the United States," *Wiley Interdisciplinary Reviews: Energy and Environment*, vol. 5, pp. 288-304, 2016.

[74] C. Joe-Wong, S. Sen, H. Sangtae, and C. Mung, "Optimized Day-Ahead Pricing for Smart Grids with Device-Specific Scheduling Flexibility," *IEEE Journal on Selected Areas in Communications,*, vol. 30, pp. 1075-1085, 2012.

[75] D. S. Kirschen, A. Rosso, M. Juan, and L. F. Ochoa, "Flexibility from the demand side," in *Power and Energy Society General Meeting, 2012 IEEE*, San Diego, CA, USA, 2012, pp. 1-6.

[76] D. Wang, S. Parkinson, W. Miao, H. Jia, C. Crawford, and N. Djilali, "Online voltage security assessment considering comfort-constrained demand response control of distributed heat pump systems," *Applied Energy*, vol. 96, pp. 104-114, 2012.

[77] P. P. Varaiya, F. F. Wu, and J. W. Bialek, "Smart operation of smart grid: Risk-limiting dispatch," *Proceedings of the IEEE*, vol. 99, pp. 40-57, 2011.

[78] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Potentials and economics of residential thermal loads providing regulation reserve," *Energy Policy*, vol. 79, pp. 115-126, 2015.

[79] B. Kladnik, A. Gubina, G. Artac, K. Nagode, and I. Kockar, "Agent-based modeling of the demand-side flexibility," in *2011 IEEE Power and Energy Society General Meeting*, Detroit, MI, USA, 2011, pp. 1-8.

[80] C. Vivekananthan, Y. Mishra, G. Ledwich, and L. Fangxing, "Demand Response for Residential Appliances via Customer Reward Scheme," *IEEE Transactions on Smart Grid*, vol. 5, pp. 809-820, 2014.

[81] B. P. Bhattarai, B. Bak-Jensen, P. Mahat, and J. R. Pillai, "Voltage controlled dynamic demand response," in *IEEE PES ISGT Europe 2013*, Copenhagen, Denmark, 2013, pp. 1-5.

[82] A. Ballanti and L. F. Ochoa, "Initial assessment of voltage-led demand response from UK residential loads," in *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2015, pp. 1-5.

[83] S. Gottwalt, J. Gärttner, H. Schmeck, and C. Weinhardt, "Modeling and valuation of residential demand flexibility for renewable energy integration," *IEEE Transactions on Smart Grid,* vol. 8, pp. 2565-2574, 2017.

[84] J. A. F. Moreno, A. M. García, A. G. Marín, E. G. Lázaro, and C. A. Bel, "An integrated tool for assessing the demand profile flexibility," *IEEE Transactions on Power Systems,* vol. 19, pp. 668-675, 2004.

[85] A. Molina, A. Gabaldon, J. Fuentes, and F. Canovas, "Approach to multivariable predictive control applications in residential HVAC direct load control," in *2000 Power Engineering Society Summer Meeting (Cat. No. 00CH37134)*, Seattle, WA, USA, 2000, pp. 1811-1816.

[86] Y. Xu and J. V. Milanović, "Day-Ahead Prediction and Shaping of Dynamic Response of Demand at Bulk Supply Points," *IEEE Transactions on Power Systems,* vol. 31, pp. 3100-3108, 2016.

[87] P.-A. Lof, T. Smed, G. Andersson, and D. Hill, "Fast calculation of a voltage stability index," *IEEE Transactions on Power Systems,* vol. 7, pp. 54-64, 1992.

[88] M. Yao, J. L. Mathieu, and D. K. Molzahn, "Using demand response to improve power system voltage stability margins," in *PowerTech, 2017 IEEE Manchester*, 2017, pp. 1-6.

[89] S. Greene, I. Dobson, and F. L. Alvarado, "Sensitivity of the loading margin to voltage collapse with respect to arbitrary parameters," *IEEE Transactions on Power Systems,* vol. 12, pp. 262-272, 1997.

[90] E. Ghahremani and I. Kamwa, "Optimal placement of multiple-type FACTS devices to maximize power system loadability using a generic graphical user interface," *IEEE Transactions on Power Systems,* vol. 28, pp. 764-778, 2013.

[91] Y. Mansour, "Voltage stability of power systems: concepts, analytical tools, and industry experience," *IEEE special publication,* 1990.

[92] "Continuation Power Flow," in *Computational Techniques for Voltage Stability Assessment and Control*, V. Ajjarapu, Ed., ed Boston, MA: Springer US, 2007, pp. 49-116.

[93] Y. Dong, X. Xie, W. Shi, B. Zhou, and Q. Jiang, "Demand-Response-Based Distributed Preventive Control to Improve Short-Term Voltage Stability," *IEEE Transactions on Smart Grid,* vol. 9, pp. 4785-4795, 2018.

[94] S. Weckx, R. D. Hulst, and J. Driesen, "Primary and Secondary Frequency Support by a Multi-Agent Demand Control System," *IEEE Transactions on Power Systems,* vol. 30, pp. 1394-1404, 2015.

[95] Y. G. Rebours, D. S. Kirschen, M. Trotignon, and S. Rossignol, "A Survey of Frequency and Voltage Control Ancillary Services&mdash;Part I: Technical Features," *IEEE Transactions on Power Systems,* vol. 22, pp. 350-357, 2007.

[96] A. Molina-Garcia, F. Bouffard, and D. S. Kirschen, "Decentralized Demand-Side Contribution to Primary Frequency Control," *IEEE Transactions on Power Systems,* vol. 26, pp. 411-419, 2011.

[97] M. Bayat, K. Sheshyekani, and A. Rezazadeh, "A Unified Framework for Participation of Responsive End-User Devices in Voltage and Frequency Control of the Smart Grid," *IEEE Transactions on Power Systems,* vol. 30, pp. 1369-1379, 2015.

[98] "Electricity Distribution Systems Losses Non-technical Overview," Ofgem, 2009, [Online].Available: https://www.ofgem.gov.uk/publications-and-updates/electricity-distribution-systems-losses-non-technical-overview.

[99] W. Hu, Z. Chen, B. Bak-Jensen, and Y. Hu, "Fuzzy adaptive particle swarm optimisation for power loss minimisation in distribution systems using optimal load response," *IET Generation, Transmission & Distribution,* vol. 8, pp. 1-10, 2014.

[100] S. Deilami, A. S. Masoum, P. S. Moses, and M. A. Masoum, "Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile," *IEEE Transactions on Smart Grid,* vol. 2, pp. 456-467, 2011.

[101] G. Heffner, C. Goldman, B. Kirby, and M. Kintner-Meyer, "Loads Providing Ancillary Services: Review of International Experience, Lawrence Berkeley National Laboratory Technical Report, LBNL-62701," *ORNL/TM-2007/060, PNNL-16618,* 2007.

[102] M. Hummon, D. Palchak, P. Denholm, J. Jorgenson, D. J. Olsen, S. Kiliccote, N. Matson, M. Sohn, C. Rose, and J. Dudley, *Grid Integration of Aggregated Demand Response: Part 2, Modeling Demand Response in a Production Cost Model*: National Renewable Energy Laboratory, 2013.

[103] D. J. Olsen, "Grid integration of aggregated demand response, part 1: load availability profiles and constraints for the western interconnection," Lawrence Berkeley National Laboratory, 2013, [Online].Available: https://cloudfront.escholarship.org/dist/prd/content/qt6ps4r3xp/qt6ps4r3xp.pdf.

[104] P. Cappers, J. MacDonald, and C. Goldman, "Market and policy barriers for demand response providing ancillary services in US markets," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States)2013.

[105] "Customer Load Active System Services," Electricity North West, 2014, [Online].Available: https://www.enwl.co.uk/globalassets/innovation/class/class-documents/class-closedown-report-master.pdf.

[106] J. Yao, W. Zhengyu, C. Jiang, and Y. Zhang, "Dispatch and bidding strategy of active distribution network in energy and ancillary services market," *Journal of Modern Power Systems and Clean Energy,* vol. 3, pp. 565-572, 2015.

[107] G. Strbac, C. K. Gan, M. Aunedi, V. Stanojevic, P. Djapic, J. Dejvises, P. Mancarella, A. Hawkes, D. Pudjianto, and S. Le Vine, "Benefits of advanced smart metering for demand response based control of distribution networks," *ENA/SEDG/Imperial College report on Benefits of Advanced Smart Metering (Version 2.0)(Energy Networks Association, London, 2010),* 2010.

[108] "Assessment of Industrial Load for Demand Response across U.S. Regions of the Western Interconnect," Oak Ridge National Laboratory, 2013.

[109] "Demand Side Flexibility Annual Report 2016 - Power Responsive," National Grid,2016, [Online].Available: http://powerresponsive.com/wp-content/uploads/2017/01/Power-Responsive-Annual-Report-2016-FINAL.pdf.

[110] J. Liang, S. K. K. Ng, G. Kendall, and J. W. M. Cheng, "Load Signature Study&#x2014;Part I: Basic Concept, Structure, and Methodology," *IEEE Transactions on Power Delivery,* vol. 25, pp. 551-560, 2010.

[111] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy and Buildings,* vol. 42, pp. 1878-1887, 2010.

[112] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR),* vol. 34, pp. 1-47, 2002.

[113] "Managing big data for smart grids and smart meters," *IBM Corporation, whitepaper (May 2012)*, 2012.

[114] Z. Dong and P. Zhang, *Emerging techniques in power system analysis*: Springer, 2010.

[115] T. Hong, D. W. Gao, T. Laing, D. Kruchten, and J. Calzada, "Training energy data scientists: universities and industry need to work together to bridge the talent gap," *IEEE Power and Energy Magazine*, vol. 16, pp. 66-73, 2018.

[116] S. M. Weiss and N. Indurkhya, *Predictive data mining: a practical guide*: Morgan Kaufmann, 1998.

[117] Z. Yiteng, O. Yew-Soon, and I. W. Tsang, "The Emerging "Big Dimensionality"," *Computational Intelligence Magazine, IEEE*, vol. 9, pp. 14-26, 2014.

[118] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*: Elsevier Science, 2005.

[119] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Elsevier, 2011.

[120] R. J. Broderick and J. R. Williams, "Clustering methodology for classifying distribution feeders," in *2013 IEEE 39th Photovoltaic Specialists Conference (PVSC)*, Tampa, FL, USA, 2013, pp. 1706-1710.

[121] H. Mori, "State-of-the-Art Overview on Data Mining in Power Systems," in *Power Systems Conference and Exposition, 2006. PSCE '06. 2006 IEEE PES*, Atlanta, GA, USA, 2006, pp. 33-34.

[122] "Matlab I., The MathWorks, Statistics and Machine Learning Toolbox User's Guide R2015b " 2015.

[123] G. K. Smyth, "Nonlinear regression," *Encyclopedia of environmetrics*, 2002.

[124] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, pp. 264-323, 1999.

[125] H. Zhao, "Decision tree technology in data classification," in *Applied Mechanics and Materials* vol. 268, ed, 2013, pp. 1752-1757.

[126] S. Kamalasadan, "Application of Artificial Intelligence Techniques in Power Systems," Asian Institute of Technology, Bangkok1998.

[127] S. S. Haykin, *Neural networks and learning machines/Simon Haykin*: New York: Prentice Hall, 2009.

[128] M. H. Beale, M. T. Hagan, and H. B. Demuth. (2017). *Matlab Neural Network Toolbox - User's Guide*.

[129] N. M. Nawi, W. H. Atomi, and M. Rehman, "The effect of data pre-processing on optimized training of artificial neural networks," *Procedia Technology*, vol. 11, pp. 32-39, 2013.

[130] T. Wang, G. Zhang, J. Zhao, Z. He, J. Wang, and M. J. Pérez-Jiménez, "Fault Diagnosis of Electric Power Systems Based on Fuzzy Reasoning Spiking Neural P Systems," *IEEE Transactions on Power Systems*, vol. 30, pp. 1182-1194, 2015.

[131] "Oracle Database Online Documentation," *[Available].Online: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#CIHHFFHB*.

[132] G. Liu, Y. Yu, F. Gao, and W. Zhu, "Research of Smart Distribution Network Big Data Model," in *23rd International Conference on Electricity Distribution, CIRED*, Lyon, France, 2015, pp. 15-18.

[133] F. De Comité, R. Gilleron, and M. Tommasi, "Learning multi-label alternating decision trees from texts and data," in *Machine Learning and Data Mining in Pattern Recognition*, ed: Springer, 2003, pp. 35-49.

[134] H. Chenchouni, T. Menasria, S. Neffar, S. Chafaa, L. Bradai, R. Chaibi, M. N. Mekahlia, D. Bendjoudi, and A. S. Bachir, "Spatiotemporal diversity, structure and trophic guilds of insect assemblages in a semi-arid Sabkha ecosystem," *PeerJ,* vol. 3, p. e860, 2015.

[135] S.-l. Yang and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews,* vol. 24, pp. 103-110, 2013.

[136] M. Craven, A. McCallum, D. PiPasquo, T. Mitchell, and D. Freitag, "Learning to extract symbolic knowledge from the World Wide Web," DTIC Document1998.

[137] M. A. Hearst, "Untangling text data mining," presented at the Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, 1999.

[138] H. Bunke, "Graph-based tools for data mining and machine learning," in *Machine Learning and Data Mining in Pattern Recognition*, ed: Springer, 2003, pp. 7-19.

[139] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine learning,* vol. 39, pp. 135-168, 2000.

[140] I. Spasić, M. Greenwood, A. Preece, N. Francis, and G. Elwyn, "FlexiTerm: a flexible term recognition method," *Journal of Biomedical Semantics,* vol. 4, pp. 1-15, 2013.

[141] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," *Introduction to Information Retrieval,* vol. 100, pp. 2-4, 2008.

[142] A. K. Singhal, "Term weighting revisited," Cornell University, 1997.

[143] Y. Seki, "Sentence Extraction by tf/idf and position weighting from Newspaper Articles," *Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering (September 2001-October 2002),* 2002.

[144] "Demand side response in the domestic sector – a literature review of major trials," Department of Energy and Climate Change, London, 2012.

[145] A. Srivastava, "Cyber Infrastructure for the Smart Grid " [Online].Available: https://sgdril.eecs.wsu.edu/files/files/Lesson%201_2_Motivation%20for%20the%20Smart%20 Grid.pdf.

[146] "Consolidated View on the ETP SG (European Technology Platform on Smart Grids) on Research, Development and Demonstration Needs in the Horizon 2020 Work Programme 2016-2017," 2015.

[147] "Guidelines for Power Quality Monitoring - Measurement Locations, Processing and Presentation of Data " CIGRE/CIRED C4.112, 2014.

[148] S. F. Noske, D.; Kolodziejczyk, K.; Helt, P., "Increase in Power Network Observability as a Data source to Improve the Efficiency of the Power Network - Results of the Pilot Smart Grid Project," 15-18 June 2015.

[149] M. Sforna, "Data mining in a power company customer database," *Electric Power Systems Research,* vol. 55, pp. 201-209, 2000.

[150] J. Alber, "State estimation in PowerFactory: Algorithmic aspects," in *Proc. RTE-VT Workshop*, 2006, pp. 1-9.

[151] "Domestic Survey Report, Customer-Led Network Revolution," Northern Powergrid (Northeast) Limited, Northern Powergrid (Yorkshire) Plc, British Gas Trading Limited, University of Durham, [Available].Online: http://www.networkrevolution.co.uk/resources/project-library2014.

[152] E. Hirst, "Barriers to Price-responsive Demand in Wholesale Electricity Markets " Edison Electric Institute Research Paper, 2002.

[153] ([Online].Available: http://www.eon.com/en/business-areas/distribution/technology-of-the-future/smart-meters/key-technology-smart-metering.html).

[154] M. Lees, "Enhanced Network Monitoring Report, Customer-led Network Revolution," Northern Powergrid (Northeast) Limited, Northern Powergrid (Yorkshire) Plc, British Gas Trading Limited, University of Durham and EA Technology Ltd., [Online].Available:http://www.networkrevolution.co.uk/resources/project-library2014.

[155] D. F. Frame, G. W. Ault, and S. Huang, "The uncertainties of probabilistic LV network analysis," in *2012 IEEE Power and Energy Society General Meeting*, San Diego, CA, USA, 2012, pp. 1-8.

[156] K. N. Hasan, M. Wang, and J. V. Milanović, "A Survey on Demand Side Management Potential in South-East Europe to Support Transmission Network Flexibility," in *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Sarajevo, Bosnia and Herzegovina, 2018, pp. 1-6.

[157] "Smart Energy GB," [Online].Available: https://www.smartenergygb.org/en/faqs.

[158] "Smart meter data - A guide to your rights and choices," [Online].Available: http://www.energy-uk.org.uk/policy/smart-meters.html.

[159] N. Jenkins, C. Long, and J. Wu, "An overview of the smart grid in Great Britain," *Engineering*, vol. 1, pp. 413-421, 2015.

[160] "The Smart Meter Report: Forecasts, regional breakdowns, costs, and savings for a top IoT device," [Available].Online: https://www.businessinsider.com/the-smart-meter-report-a-look-at-how-the-smart-meter-market-is-evolving-2015-3?r=US&IR=T.

[161] "Smart Metering Equipment Tecnical Specifications," [Online].Available https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/299395/smets.pdf.

[162] [Online].Available: http://www.rexelenergysolutions.co.uk/media/pdfs/EL09_AS230_Single_Phase_Smart_Meter_Technical.pdf

[163] *[Online]*. Available: http://www.smsmetering.co.uk/wp-content/uploads/2013/08/ME372-Data-Sheet_SMS.pdf

[164] ([Online].Available: http://www.jwsmartmeters.co.uk/Landis-Gyr-E470).

[165] N. Uribe-Pérez, L. Hernández, D. de la Vega, and I. Angulo, "State of the art and trends review of smart metering in electricity grids," *Applied Sciences*, vol. 6, p. 68, 2016.

[166] "Cost-benefit analyses & state of play of smart metering deployment in the EU-27," European Commission, 2014.

[167] L. Thomas and N. Jenkins, "Smart metering for the UK," HubNet Position Paper Series, 2012.

[168] "Introduction to the UK Smart Grid Sector," [Online].Available: https://www.techuk.org/.

[169] A. Al-Wakeel, J. Wu, and N. Jenkins, "K-means based load estimation of domestic smart meter measurements," *Applied Energy,* vol. 194, pp. 333-342, 2017.

[170] C. G. Zhao, C.; Li, F., "Classification of Low Voltage Distribution Networks Based on Fixed Data," in *23rd Inernational Conference on Electricity Distribution (CIRED)*, Lyon, France, 2015.

[171] N. D. Hatziargyriou, A.; Korres, N.; Dova, F.;Gkavogianni, A.; Vlachos,Y.; Koukoula, D.; Tsitsimelis, A. , "A Data Repository for Automated Evaluation of Smart Grid Solutions," in *23rd International Conference on Electricity Distribution (CIRED)*, Lyon, France, 2015.

[172] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, "Big data analytics in power distribution systems," in *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)* 2015, pp. 1-5.

[173] H. Dong, M. Jin, R.-m. He, and D. Zhao-yang, "A Real Application of Measurement-Based Load Modeling in Large-Scale Power Grids and its Validation," *IEEE Transactions on Power Systems,* vol. 24, pp. 1756-1764, 2009.

[174] S. Tso, J. Lin, H. Ho, C. Mak, K. Yung, and Y. Ho, "Data mining for detection of sensitive buses and influential buses in a power system subjected to disturbances," *IEEE Transactions on Power Systems*, vol. 19, pp. 563-568, 2004.

[175] Y. R. Gahrooei, A. Khodabakhshian, and R.-A. Hooshmand, "A New Pseudo Load Profile Determination Approach in Low Voltage Distribution Networks," *IEEE Transactions on Power Systems,* vol. 33, pp. 463-472, 2018.

[176] A. Moraru and D. Mladenić, "Complex event processing and data mining for smart cities," in *Conference on Data Mining and Data Warehouses (SkiDD 2013), Held at the 15th International Multiconference on Information Society (IS-2012), 8th October*, Ljubljana, Slovenia, 2012.

[177] T. Guo, P. Papadopoulos, P. Mohammed, and J. V. Milanovic, "Comparison of ensemble decision tree methods for on-line identification of power system dynamic signature considering availability of PMU measurements," in *PowerTech, 2015 IEEE Eindhoven*, 2015, pp. 1-6.

[178] A. Capozzoli, M. S. Piscitelli, S. Brandi, D. Grassi, and G. Chicco, "Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings," *Energy,* vol. 157, pp. 336-352, 2018.

[179] D. C. Park, M. A. El-Sharkawi, R. J. Marks, II, L. E. Atlas, and M. J. Damborg, "Electric load forecasting using an artificial neural network," *IEEE Transactions on Power Systems*, vol. 6, pp. 442-449, 1991.

[180] I. P. Panapakidis, G. K. Papagiannis, and G. C. Christoforidis, "Bus load forecasting via a combination of machine learning algorithms," in *Power Engineering Conference (UPEC), 2014 49th International Universities*, Cluj-Napoca, Romania, 2014, pp. 1-6.

[181] S. V. Verdú, M. O. Garcia, C. Senabre, A. G. Marín, and F. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Transactions on Power Systems,* vol. 21, pp. 1672-1682, 2006.

[182] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, and M. Cordeiro, "A comparative analysis of clustering algorithms applied to load profiling," in *Machine Learning and Data Mining in Pattern Recognition*, ed: Springer, 2003, pp. 73-85.

[183] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting—A novel pooling deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, pp. 5271-5280, 2018.

[184] H. Liao, J. V. Milanović, M. Rodrigues, and A. Shenfield, "Voltage Sag Estimation in Sparsely Monitored Power Systems Based on Deep Learning and System Area Mapping," *IEEE Transactions on Power Delivery,* vol. 33, pp. 3162-3172, 2018.

[185] "Quality of Supply and Market Regulation; Survey within Europe," KEMA Consulting2006.

[186] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.

[187] R. Preece, N. C. Woolley, and J. V. Milanovic, "NaFIRS Data Analysis - Correlation between quality of supply and network performance data," Internal report, School of Electrical and electronic Engineering, The University of Manchester, 2012.

[188] "The Value of Lost Load (VoLL) for Electricity in Great Britain, final report for OFGEM and DECC," London Economics2013, [Online].Available: https://www.ofgem.gov.uk/ofgem-publications/82293/london-economics-value-lost-load-electricity-gbpdf.

[189] "VirginiaTech Research Data," [Online].Available: http://www.ari.vt.edu/research-data/.

[190] J. Hu, J. Cao, M. Z. Q. Chen, J. Yu, J. Yao, S. Yang, and T. Yong, "Load Following of Multiple Heterogeneous TCL Aggregators by Centralized Control," *IEEE Transactions on Power Systems*, vol. 32, pp. 3157-3167, 2017.

[191] X. Chen, C. Kang, X. Tong, Q. Xia, and J. Yang, "Improving the Accuracy of Bus Load Forecasting by a Two-Stage Bad Data Identification Method," *IEEE Transactions on Power Systems*, vol. 29, pp. 1634-1641, 2014.

[192] "D3.4 Smart meters architecture and data model analysis ", NOBEL GRID project, [Online].Available: http://nobelgrid.eu/deliverables/ , 2016.

[193] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association,* vol. 74, pp. 829-836, 1979.

[194] "The impact of changing energy use patterns in buildings on peak electricity demand in the UK," Building Research Establishment Ltd, 2008.

[195] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting,* vol. 22, pp. 679-688, 2006.

[196] Y. Xu and J. V. Milanović, "Accuracy of ANN based methodology for load composition forecasting at bulk supply buses," in *2014 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Durham, UK, 2014, pp. 1-6.

[197] F. Lamberti, D. Cuicai, V. Calderaro, and L. F. Ochoa, "Estimating the load response to voltage changes at UK primary substations," in *2013 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Copenhagen, Denmark, 2013, pp. 1-5.

[198] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Medical Journal,* vol. 24, pp. 69-71, 2012.

[199] M. Nijhuis, "Long-term Planning of Low Voltage Networks," PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2017.

[200] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," *IEEE Transactions on Power Systems,* vol. 16, pp. 44-55, 2001.

[201] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*: John Wiley & Sons, 2015.

[202] J. W. Taylor and P. E. McSharry, "Short-Term Load Forecasting Methods: An Evaluation Based on European Data," *IEEE Transactions on Power Systems*, vol. 22, pp. 2213-2219, 2007.

[203] (2017). *Matlab Econometrics Toolbox User's Guide R2017b, The MatWorks*

[204] P. Han, P. X. Wang, S. Y. Zhang, and D. H. Zhu, "Drought forecasting based on the remote sensing data using ARIMA models," *Mathematical and computer modelling,* vol. 51, pp. 1398-1403, 2010.

[205] H. Kun-Yuan and H. Yann-Chang, "Integrating direct load control with interruptible load management to provide instantaneous reserves for ancillary services," *IEEE Transactions on Power Systems,* vol. 19, pp. 1626-1634, 2004.

[206] W. Deh-Chang and C. Nanming, "Air conditioner direct load control by multi-pass dynamic programming," *IEEE Transactions on Power Systems,* vol. 10, pp. 307-313, 1995.

[207] K. M. Kosa, S. C. Cates, S. L. Godwin, R. J. Coppings, and L. Speller-Henderson, "Most Americans are not prepared to ensure food safety during power outages and other emergencies," *Food Protection Trends,* vol. 31, pp. 428-436, 2011.

[208] F. Milano, *Power system modelling and scripting*: Springer Science & Business Media, 2010.

[209] Z. Bo and C. Yi-Jia, "Multiple objective particle swarm optimization technique for economic load dispatch," *Journal of Zhejiang University-Science A,* vol. 6, pp. 420-427, 2005.

[210] T. Niknam, M. Narimani, J. Aghaei, and R. Azizipanah-Abarghooee, "Improved particle swarm optimisation for multi-objective optimal power flow considering the cost, loss, emission and voltage stability index," *IET Generation, Transmission & Distribution,* vol. 6, pp. 515-527, 2012.

[211] R. Hassan, B. Cohanim, O. De Weck, and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm," in *46th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference*, Austin, TX, USA, 2005, p. 1897.

[212] R. D. Zimmerman and C. E. Murillo-Sánchez, "Matpower 6.0 User's Manual," ed: December, 2016.

[213] R. C. Eberhart and Y. Shi, "Comparison between genetic algorithms and particle swarm optimization," in *International conference on evolutionary programming*, San Diego, CA, USA, 1998, pp. 611-616.

[214] D. W. Boeringer and D. H. Werner, "Particle swarm optimization versus genetic algorithms for phased array synthesis," *IEEE Transactions on antennas and propagation,* vol. 52, pp. 771-779, 2004.

[215] M. Clerc, *Particle swarm optimization* vol. 93: John Wiley & Sons, 2010.

[216] J. V. Paatero and P. D. Lund, "A model for generating household electricity load profiles," *International journal of energy research,* vol. 30, pp. 273-290, 2006.

[217] J. Jardini, C. Tahan, S. Ahn, and E. Ferrari, "Distribution transformer loading evaluation based on load profiles measurements," *IEEE Transactions on Power Delivery,* vol. 12, pp. 1766-1770, 1997.

[218] J. Silvente, G. M. Kopanos, E. N. Pistikopoulos, and A. Espuña, "A rolling horizon optimization framework for the simultaneous energy supply and demand planning in microgrids," *Applied Energy,* vol. 155, pp. 485-501, 2015.

Appendix A: IEEE 33-bus network model data

Table A 1 Bus data

Bus number	Bus type	Real power (MW)	Reactive power (Mvar)	Voltage (kV)
1	PV	0	0	12.66
2	PQ	0.1	0.06	12.66
3	PQ	0.09	0.04	12.66
4	PQ	0.12	0.08	12.66
5	PQ	0.06	0.03	12.66
6	PQ	0.06	0.02	12.66
7	PQ	0.2	0.1	12.66
8	PQ	0.2	0.1	12.66
9	PQ	0.06	0.02	12.66
10	PQ	0.06	0.02	12.66
11	PQ	0.045	0.03	12.66
12	PQ	0.06	0.035	12.66
13	PQ	0.06	0.035	12.66
14	PQ	0.12	0.08	12.66
15	PQ	0.06	0.01	12.66
16	PQ	0.06	0.02	12.66
17	PQ	0.06	0.02	12.66
18	PQ	0.09	0.04	12.66
19	PQ	0.09	0.04	12.66
20	PQ	0.09	0.04	12.66
21	PQ	0.09	0.04	12.66
22	PQ	0.09	0.04	12.66
23	PQ	0.09	0.05	12.66
24	PQ	0.42	0.2	12.66
25	PQ	0.42	0.2	12.66
26	PQ	0.06	0.025	12.66
27	PQ	0.06	0.025	12.66
28	PQ	0.06	0.02	12.66
29	PQ	0.12	0.07	12.66
30	PQ	0.2	0.1	12.66
31	PQ	0.15	0.07	12.66
32	PQ	0.21	0.1	12.66
33	PQ	0.06	0.04	12.66
34	SLACK	0	0	12.66

Table A 2 Branch data

From bus	To bus	Resistance (Ω)	Reactance (Ω)
1	2	0.057525912	0.029324
2	3	0.307595167	0.156668
3	4	0.228356656	0.1163
4	5	0.237777928	0.121104
5	6	0.510994811	0.441115
6	7	0.116798814	0.386085
7	8	0.44386045	0.146685
8	9	0.642643047	0.461705
9	10	0.651378001	0.461705
10	11	0.122663712	0.040555
11	12	0.233597628	0.077242
12	13	0.915922324	0.720634
13	14	0.337917936	0.444796
14	15	0.368739846	0.328185
15	16	0.465635443	0.340039
16	17	0.804239697	1.073775
17	18	0.456713311	0.358133
19	20	0.938508419	0.845668
20	21	0.255497406	0.298486
21	22	0.442300637	0.584805
3	23	0.28151509	0.192356
23	24	0.560284909	0.442425
24	25	0.559037059	0.437434
6	26	0.126656834	0.064514
26	27	0.177319567	0.090282
27	28	0.660736881	0.582559
28	29	0.501760717	0.437122
29	30	0.316642084	0.161285
30	31	0.607952801	0.60084
31	32	0.193728802	0.225799
32	33	0.212758523	0.330805
2	34	0.057525912	0.029324
34	19	0.938508419	0.845668

Table A 3 Generator data

Bus number	Real power (MW)	Reactive power (Mvar)	Real power limits MIN/MAX (MW)	Reactive power limits MIN/MAX (Mvar)	Cost function
1	1.851	0	1.851/1.851	-5/5	$C_g = 10 \cdot P_g$
34	0	0	-5/5	-5/5	$C_g = 1000 \cdot {P_g}^2$

Table A 4 and Figure A 1 show results of the power flow run on IEEE 33 bus network in Matpower and DIgSILENT/PowerFactory. Table A 4 represents the general results (generation outputs and losses), while Figure A 1 illustrates the difference between bus voltages. Matching results prove the validity of using these two types of software in simulations run for the analyses described in Chapter 4.

Table A 4 General power flow results

	Matpower (MW/Mvar)	DIgSILENT/PowerFactory (MW/Mvar)
Generation	1.85/0.69	1.85/0.69
Grid infeed from slack bus	2.02/1.22	2.02/1.23
Total losses	0.16/0.12	0.16/0.11

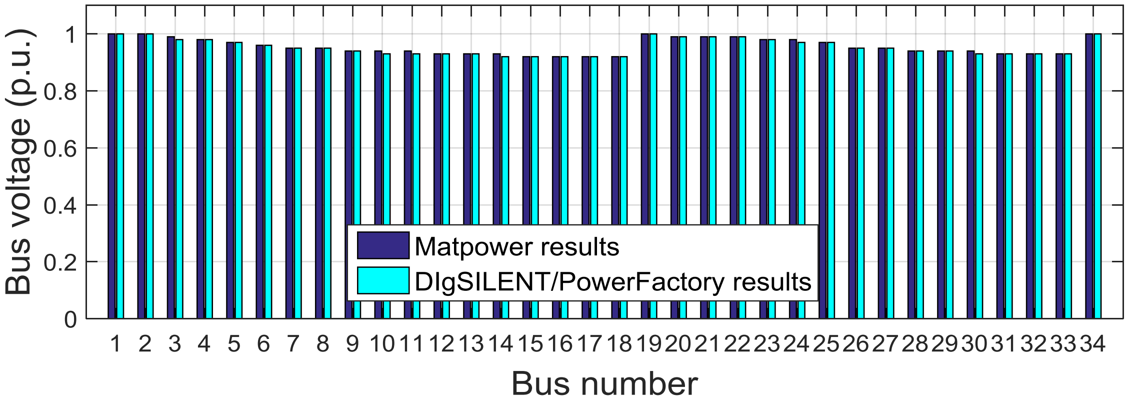


Figure A 1 Comparison of power flow results with respect to bus voltages

Appendix B: List of author’s thesis based publications

B 1. Journal papers

Published journal papers:

B1. J. Ponoćko and J. V. Milanović, "Forecasting Demand Flexibility of Aggregated Residential Load Using Smart Meter Data," *IEEE Transactions on Power Systems*, vol. 33, pp. 5446-5455, 2018 (DOI: 10.1109/TPWRS.2018.2799903)

Submitted journal papers:

B2. J. Ponoćko and J. V. Milanović ,”Multi-objective Demand Side Management at Distribution Network Level in Support of Transmission Network Operation", submitted to the IEEE Transactions on Power Systems, 2018 (under 2nd review)

B 2. Conference papers

Published conference papers:

B3. J. Ponocko, J. V. Milanović, R. Preece and N. C. Woolley, "Application of data analytics for information retrieval from a typical DSO's database," *2016 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Ljubljana, 2016, pp. 1-6.

B4. J. Ponocko and J. V. Milanovic, "Comparative analysis of data availability and data requirements for efficient management and control of future distribution networks," *Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower 2016)*, Belgrade, 2016, pp. 1-7

B5. K. Li, J. Ponocko, L. Zhang and J. V. Milanovic, "Methodology for close to real time profiling of aggregated demand using data streams from smart meters," *Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower 2016)*, Belgrade, 2016, pp. 1-8

B6. P. Chen, J. Ponocko, N. Milosevic, G. Nenadic and J. V. Milanovic, "Towards application of text mining for enhanced power network data analytics — Part I: Retrieval and ranking of textual data from the internet," *Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower 2016)*, Belgrade, 2016, pp. 1-8.

B7. Y. Chen, J. Ponocko, N. Milosevic, G. Nenadic and J. V. Milanovic, "Towards application of text mining for enhanced power network data analytics — Part II: Offline analysis of textual data," *Mediterranean Conference on Power*

Generation, Transmission, Distribution and Energy Conversion (MedPower 2016), Belgrade, 2016, pp. 1-8.

B8. J. Ponoćko and J. V. Milanović, "Smart Meter-Driven Estimation of Residential Load Flexibility," in CIRED conference, Glasgow, UK, 2017

B9. J. Ponoćko and J. V. Milanović, "Application of data analytics for advanced demand profiling of residential load using smart meter data," in *2017 IEEE Manchester PowerTech*, 2017, pp. 1-6.

B10. J. Ponocko, J. Cai, Y. Sun, and J. V. Milanovic, "Real-time visualisation of residential load flexibility for advanced demand side management," in *2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON)*, 2018, pp. 181-186.

B11. J. Ponoćko and J. V. Milanović, "Towards the Advanced Demand Response in Distribution Network", in the XI Conference on Electricity Distribution of Serbia with Regional Participation (CIRED Serbia), Kopaonik, Serbia, September 24-28, 2018

B12. J. Ponoćko and J. V. Milanović, "Data Requirements for a Reliable Demand Decomposition in Sparsely Monitored Power Networks," in *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 2018, pp. 1-6.

B13. J. Ponoćko and J. V. Milanović, "The Effect of Load-follow-generation Motivated DSM Programme on Losses and Loadability of a Distribution Network with Renewable Generation", in IEEE PES GTD Grand International Conference and Exposition Asia (GTDA) 2019, Bangkok, Thailand, 2019

B14. M. Wang, J. Ponoćko and J. V. Milanović, "The Effect of the Type and Composition of Demand on the Influence of DSM on Power System Angular Stability", in IEEE PES GTD Grand International Conference and Exposition Asia (GTDA) 2019, Bangkok, Thailand, 2019

B 3. Technical reports

Though produced while working on the thesis, some of the reports and their content are not directly related to the research presented in this thesis:

B15. Alma Solar, Rafael Alemany, Vicente Ortuno, Abel Ribes, Huilian Liao, Jelena Ponoćko, Jovica V. Milanović, et al., Distribution grid and retail market Requirements definition, Deliverable 1.1, EU HORIZON 2020 project NOBEL GRID, July 2015

B16. Alma Solar, Rafael Alemany, Vicente Ortuno, Abel Ribes, Huilian Liao, Jelena Ponoćko, Jovica V. Milanović, et al., Distribution grid and retail market Scenarios and Use Case definition, Deliverable 1.2, EU HORIZON 2020 project NOBEL GRID, July 2015

B17. Paul Smith, Ewa Piatkowska, Jelena Ponoćko, Huilian Liao, et al., Smart grids reference architecture and data models v1, Deliverable 3.1, EU HORIZON 2020 project NOBEL GRID, January 2016

B18. Paul Smith, Ewa Piatkowska, Jelena Ponoćko, Xiaoqing Tang, et al., Smart grids reference architecture and data models v2, Deliverable 3.2, EU HORIZON 2020 project NOBEL GRID, January 2017

B19. Joel Höglund, Joakim Ericsson, Mihai Sanduleac, Lucas Pons, Lola Alacreu, Papapolyzos Thomas, Aleksandar Hudic, Thomas Hecht, Xiaoqing Tang, Jelena Ponoćko, Smart meters architecture and data model analysis v2, Deliverable 3.4, EU HORIZON 2020 project NOBEL GRID, January 2017

B20. Jan Ringelstein, Jelena Ponoćko, Xiaoqing Tang, Jovica V. Milanović, et al., Integration and Lab-testing Models for distributed generation and storage integration, Deliverable 7.5, EU HORIZON 2020 project NOBEL GRID, August 2017

B21. Jonathan Atkinson, Ben Aylott, Jelena Ponoćko, Xiaoqing Tang, Joel Höglund, Jovica V. Milanović, Lola Alacreu Garcia, Diego García-Casarrubios Gálvez, Ex-ante analysis Manchester pilot site, Deliverable 15.1, EU HORIZON 2020 project NOBEL GRID, November 2017

B22. Lola Alacreu, M Carmen Bueno, Jelena Ponoćko, et al., Exploitation activities period 1, Deliverable 22.1, EU HORIZON 2020 project NOBEL GRID, June 2016

B23. Lola Alacreu, M Carmen Bueno, Jelena Ponoćko, et al., Exploitation activities period 2, Deliverable 22.2, EU HORIZON 2020 project NOBEL GRID, June 2017

B24. Kazi Hasan, Mathaios Panteli, Alessandra Parisio, Xiaoqing Tang, Jelena Ponoćko, Jovica V. Milanović, et al., "CROSSBOW project requirements definition" Deliverable 2.1, EU H2020 Project "CROSS BOrder management of variable renewable energies and storage units enabling a transnational Wholesale market (CROSSBOW)", (H2020-773430), August 2018

B25. Jelena Ponoćko, Kazi Hasan, Alessandra Parisio, Mathaios Panteli, Xiaoqing Tang, Jovica V. Milanović, et al., "CROSSBOW Use cases, scenarios and KPIs identification" Deliverable 2.2, EU H2020 Project "CROSS BOrder management of variable renewable energies and storage units enabling a transnational Wholesale market (CROSSBOW)", (H2020-773430), August 2018

B26. Alessandra Parisio, Mathaios Panteli, Jelena Ponoćko, Kazi Hasan, Mengxuan Wang, Jovica V. Milanović, et al., "CROSSBOW demo clusters formal analysis" Deliverable 2.3, EU H2020 Project "CROSS BOrder management of variable renewable energies and storage units enabling a transnational Wholesale market (CROSSBOW)", (H2020-773430), August 2018