DISS. ETH NO. 27435

DATA ANALYTICS AND MACHINE LEARNING FOR THE OPERATION AND PLANNING OF DISTRIBUTION GRIDS

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

THIERRY ZUFFEREY

MSc D-ITET ETH Zurich

born on 27 June 1991

citizen of
Noble-Contrée (VS), Switzerland

accepted on the recommendation of

Prof. Dr. Gabriela Hug, examiner
Prof. Dr. Gustavo Valverde, co-examiner
Prof. Dr. Pierre Pinson, co-examiner

2021

ETH Zurich
EEH - Power Systems Laboratory
ETL G22
Physikstrasse 3
8092 Zurich, Switzerland

DOI: 10.3929/ethz-b-000488631
ISBN: 978-3-907234-48-8

© Thierry Zufferey, 2021
For a copy visit: https://www.eeh.ee.ethz.ch

Printed in Switzerland by the ETH Druckzentrum

To my family and friends.

ABSTRACT

The recent aspirations for a more sustainable energy system and a reduction of energy-related CO2 emissions have triggered a change of paradigm in power distribution grids, often encouraged by national and supranational policies. Traditionally considered a passive black-box component of power systems, the distribution grid currently undergoes a rapid transformation and sees the emergence of new types of loads (e.g., electric vehicles, electric heating systems, electric water heaters) as well as distributed energy resources (e.g., small wind turbines, solar photovoltaic systems, battery energy storage systems). Their integration requires increased reliability, efficiency, and adaptability of distribution systems, which inevitably relies on more visibility. Consequently, advanced electricity sensor elements are massively rolled out in distribution grids down to the end-users. The gains in transparency and controllability offered by the advanced metering infrastructure open up a wide range of new opportunities discussed extensively in the literature. Nevertheless, the research community is usually not granted access to real-world data due to understandable privacy concerns. It must depend on simplifications and synthetic data that often do not reflect the more complex reality and might lead to biased conclusions. On the sole basis of real-world data, this thesis intends to highlight which are the assumptions that can realistically be taken in the development and validation of data-based studies and applications. It also suggests various processes and methods to effectively leverage the actual potential of the advanced metering infrastructure and address some of the current challenges in grid operation and planning. This work primarily focuses on the low-voltage level, which is still rarely considered in the state-of-the-art literature. Data preparation, big data visualization, pseudo-measurement synthesis, distribution system state estimation, load disaggregation, and short-term forecasting are among the investigated topics. In that respect, the thesis hopes to bridge some of the gaps between the relatively conservative practices in the power industry and the various advanced data-based applications proposed in the literature.

RÉSUMÉ

Les récentes aspirations à un système énergétique plus durable et à une réduction des émissions de CO2 liées à la production énergétique ont provoqués un changement de paradigme dans le réseau de distribution électrique, souvent encouragé par des politiques nationales et supranationales. Traditionnellement considéré comme un élément passif du système électrique, le réseau de distribution subit actuellement une transformation rapide. De nouveaux acteurs émergent, tels que les véhicules électriques, les systèmes de chauffage électrique, ou les chauffe-eau électriques, mais aussi de nouvelles ressources énergétiques à petite échelle comme les éoliennes domestiques, les panneaux solaires photovoltaïques, ou les batteries domestiques. Leur intégration nécessite une fiabilité, une efficacité et une adaptabilité accrues des réseaux de distribution, ce qui nécessite inévitablement une plus grande visibilité. Par conséquent, des compteurs électriques communicants de nouvelle génération sont massivement déployés dans les réseaux de distribution jusqu'aux utilisateurs finaux. Les gains en transparence et en contrôlabilité offerts par cette infrastructure de mesurage avancée ouvrent un large éventail de nouvelles opportunités qui sont largement discutées dans la littérature. Néanmoins, la communauté scientifique n'a généralement pas accès aux données réelles en raison de préoccupations compréhensibles en matière de confidentialité. Elle doit donc s'appuyer sur des simplifications et des données synthétiques qui, souvent, ne reflètent pas la réalité plus complexe et peuvent conduire à des conclusions biaisées. Sur la seule base de données réelles, cette thèse vise à mettre en évidence les hypothèses qui peuvent être prises de manière réaliste dans le développement et la validation des études et des applications se nourrissant de données. Elle suggère également divers processus et méthodes permettant d'exploiter efficacement le potentiel réel de l'infrastructure de mesurage avancée et de relever certains des défis actuels en matière de gestion et de planification des réseaux de distributions. Ce travail se concentre tout particulièrement sur le réseau basse tension, auquel sont connectés la majorité des utilisateurs finaux, et qui est encore rarement pris en compte dans la littérature actuelle. La préparation des données, leur visualisation, la création de données synthétiques, l'estimation de l'état du réseau de distribution, la désagrégation de profils de charge et les prévisions à court terme font partie des sujets étudiés. À cet égard, la thèse espère combler une partie du fossé qui

peut être observé entre les pratiques relativement conservatrices de l’industrie électrique et les diverses applications avancées proposées dans la littérature sur la base des données.

ACKNOWLEDGEMENTS

This doctoral thesis summarizes the research activities conducted during my time as a PhD student at the Power Systems Laboratory (PSL) of ETH Zurich from May 2016 to April 2021. There are many people to thank for their help and support during this time.

First of all, my deepest gratitude goes to Prof. Gabriela Hug for accepting me as a part of PSL and giving me the opportunity to pursue a PhD in her lab. I highly appreciate her support, time, understanding, trust, and recognition of my work. I also feel fortunate for the great freedom of research as well as the multiple opportunities for self-improvement and continuous training around the world. Her very good tact and diplomacy have additionally led to solving efficiently conflicting situations. Besides, Gaby particularly values the notions of respect and work-life balance and managed to build a working environment that gave me intrinsic motivation.

The story of my PhD starts with Adaptricity. After completing my Master thesis in this newly founded start-up company, I got encouraged by Stephan Koch and especially Andreas Ulbig to carry on with a PhD while keeping a foot in their team. I'm sincerely grateful for this unique opportunity, which has profoundly shaped my vision of research towards real-world applications. It was also a privilege to be a direct witness of the many achievements, but also difficulties, of a start-up company at the early stage of its development. In this context, I would also like to express my gratitude to my former colleagues at Adaptricity, especially Diren Toprak, Damiano Toffanin, and Nicolas Stocker, for the inspiring discussions.

Moreover, this thesis would not be the same without the enormous work performed by the multiple students I had the chance to supervise during my PhD studies. Beyond the sole completion of a semester or Master thesis, they have considerably enriched my understanding of the different investigated topics. In chronological order, I would like to thank Alice Lépouzé, David Rodriguez Flores, Joël Dunant, Carmen Exner, Benjamin Schaule, Stefan Agovski, Roman Engeler, Haoyang Zhang, Iraklis Katsolas, Iason-Ioannis Chontzopoulos, Sandro Renggli, Filippo Ferrando, Jochen Stiasny, Tara Katamay-Smith, Julie Rousseau, Leandro Von Krannichfeldt, Shipra Mohan, Timmy Frischknecht, and Nami Hekmat. I wish them all the best for their professional career, sometimes in academia, often in the industry.

The very friendly and supportive atmosphere within PSL has also greatly contributed to my PhD experience, being at the lab or during the multiple conferences, excursions, and ski weekends. My appreciation generally goes to each of my current and former colleagues. I specifically would like to thank Jo, Andrew, Olivier, and Conleigh for the many serious and amusing discussions in G22, Jack for the often late brainstorming sessions at the door between our respective offices, Conor and Adrian for the co-organisation of multiple events, Line followed by Johanna for being so motivating for the PSL running of Thursday, Stavros for his delicious Greek ice coffees, and lastly Judith for the various talks *uf Schwiizerdütsch* about everything and anything.

Toward the end of my PhD studies, I had the opportunity to spend five awesome months at the University of Costa Rica, under the supervision of Prof. Gustavo Valverde. I am deeply grateful for his great guidance, feedback, support, and hospitality. On the one hand, the stay allowed me to get a new perspective on my research, where I notably got the chance to present my work in front of a broad audience with system operators, smart meter manufacturers, and policy makers. On the other hand, this will remain an unforgettable experience on a personal level thanks to the numerous people who introduced me to the *tico* lifestyle. I also would like to thank Gustavo for being a co-examiner of this thesis and for challenging me with fundamental questions during the defense.

Another special thanks goes to Prof. Pierre Pinson for accepting to be on my doctoral committee and for providing valuable feedback that helped me improve the quality of the final manuscript.

Besides, I was particularly lucky to principally work on real-world data for the purpose of my PhD, especially via the intermediary of Adaptricity and Gustavo Valverde. In that respect, I am thankful to Industrielle Werke Basel, and particularly Dirk Schmidt, Compañía Nacional de Fuerza y Luz, and Jairo Quirós-Tortós for granting me access to their valuable data sets.

Last but not least, I would not be at this stage without the precious support and love from my family and my friends. They have always encouraged my choices and been on my side in both good and bad times. I sincerely want to tell them a big MERCI. In this context, Johanna definitely deserves a special place. Her joy of life and positive energy have been a true breath of fresh air, especially during the most stressful periods of my PhD.

Thierry Zufferey

August 2021

CONTENTS

List of Acronyms xv
List of Figures xx
List of Tables xxviii
1 INTRODUCTION 1
  1.1 Background and Motivation 1
  1.2 Contributions 3
  1.3 Thesis Outline 4
  1.4 Publications 6

I DATA IN DISTRIBUTION GRIDS

2 DIGITALIZATION OF DISTRIBUTION GRIDS 13
  2.1 Introduction 13
  2.2 Advanced Metering Infrastructure 18
    2.2.1 Smart Meters 18
    2.2.2 Communication Networks 19
    2.2.3 Meter Data Management System 21
  2.3 Roll-Out of Smart Electricity Meters 22
    2.3.1 Situation in Switzerland 22
    2.3.2 Situation in the European Union 24
    2.3.3 Situation in the World 27
  2.4 Smart Meter Data Applications and Challenges 29
    2.4.1 Detection of Non-Technical Losses 30
    2.4.2 Monitoring and Situational Awareness 31
    2.4.3 Energy Forecasting 34
    2.4.4 Demand-Side Management 36
    2.4.5 Transactive Energy Systems 40
    2.4.6 Customer Concerns and Data Privacy 41
  2.5 Swiss Start-Up Companies and Smart Grid Solutions 44
    2.5.1 Adaptricity 45
    2.5.2 Depsys 47
    2.5.3 Exnaton 49
  2.6 Conclusion 50
3 DATA SETS AND PREPARATION 53
  3.1 Introduction 53

3.2 Presentation of AMI Data Sets 54
3.2.1 Distribution System of the City of Basel 58
3.2.2 Distribution System in Costa Rica 62
3.2.3 Sub-Metering Study in Costa Rica 63
3.3 Data Preparation 67
3.3.1 Data Standardization and Formatting 71
3.3.2 Statistical Analysis 77
3.3.3 Outlier Detection 80
3.3.4 Anomaly Detection, Filtering and Cleaning 82
3.3.5 Missing Data Imputation 87
3.3.6 Data Validation and Fusion 92
3.4 Conclusion 95
4 SMART METER DATA CLUSTERING AND VISUALIZATION 99
4.1 Introduction 99
4.2 K-Means Clustering Algorithm 101
4.3 Data Preparation and Features Extraction 104
4.4 Visualization of Clustering Outcome 106
4.4.1 Temporal representation 107
4.4.2 Spatial Representation 111
4.5 Conclusion 115
5 TEMPORAL RESOLUTION AND SPATIAL AGGREGATION 119
5.1 Introduction 119
5.2 Data Preparation 121
5.2.1 Costa Rican Sub-Metering Data 122
5.2.2 Costa Rican Smart Meter Data 122
5.3 Effect of Temporal Resolution 123
5.3.1 Individual Electrical Appliances 123
5.3.2 Individual Residential End-Users 127
5.3.3 Aggregation of Residential End-Users 129
5.4 Combined Effect of Spatial and Temporal Dimensions 131
5.5 Conclusion 136

II DATA-BASED MODELING OF DISTRIBUTION GRIDS

6 PSEUDO-MEASUREMENT SYNTHESIS 141
6.1 Introduction 141
6.2 Methodology 143
6.2.1 Power Gap Profile 145
6.2.2 Synthesis of Active Power Profiles 146
6.2.3 Synthesis of Reactive Power Profiles 153

6.2.4 Optimal Aggregate Matching 160
6.3 Case Studies 163
6.3.1 Active Power Pseudo-Measurements 163
6.3.2 Reactive Power Pseudo-Measurements 169
6.4 Conclusion 173
7 IMPACT OF DATA ON DISTRIBUTION SYSTEM STATE ESTIMATION 175
7.1 Introduction 175
7.2 Distribution System State Estimation 178
7.2.1 Weighted Least Squares Algorithm 178
7.2.2 Measurements and Standard Deviation 178
7.3 Setup of Sensitivity Analysis 182
7.3.1 Benchmark Grid Model 183
7.3.2 Dimensions 185
7.3.3 Evaluation Metrics 191
7.4 Results of Sensitivity Analysis 193
7.4.1 Power Injection 194
7.4.2 Component Loading 197
7.4.3 Voltage 197
7.5 Computational Complexity and Scalability 200
7.6 Conclusion 201

III APPLICATION OF DATA-BASED APPROACHES IN LOW-VOLTAGE GRIDS

8 DETECTION AND DISAGGREGATION OF FLEXIBLE RESIDENTIAL LOADS 207
8.1 Introduction 207
8.2 Methodology 211
8.2.1 Detection of Cold Appliance Load 212
8.2.2 Detection of Water Heater Load 213
8.2.3 Disaggregation of Cold Appliance Load 215
8.2.4 Disaggregation of Water Heater Load 220
8.3 Performance Evaluation 222
8.3.1 Data Set 223
8.3.2 Evaluation Metrics 224
8.3.3 Disaggregation of Cold Appliance Load 226
8.3.4 Disaggregation of Water Heater Load 232
8.4 Share of Flexible Loads at an Aggregate Level 235
8.5 Conclusion 236

9 DETERMINISTIC AND PROBABILISTIC SHORT-TERM FORECASTING 239
9.1 Introduction 239
9.2 Forecasting Workflow 242
9.3 Inadequacy of Standard Deterministic Approaches 244
9.3.1 Forecasting Algorithms 246
9.3.2 Deterministic Evaluation Metrics 249
9.3.3 Statistical Properties 253
9.3.4 Performance Evaluation 256
9.4 Short-term Probabilistic State Forecasting 260
9.4.1 Methodology 261
9.4.2 Performance Evaluation 265
9.5 Preventive Voltage Control under Uncertainty 275
9.5.1 Methodology 277
9.5.2 Results and Discussion 281
9.6 Conclusion 284
10 CONCLUSIONS AND OUTLOOK 287
10.1 Summary 287
10.2 Conclusions 290
10.3 Outlook 292

BIBLIOGRAPHY 295

LIST OF ACRONYMS

AC	Air Conditioning
ACE	Average Coverage Error
ADT	Anomaly Detection Technique
AIC	Akaike Information Criterion
AIS	Average Interval Score
AMCM	Adaptive Markov Chain Model
AMI	Advanced Metering Infrastructure
AMR	Automated Meter Reading
ANN	Artificial Neural Network
AR	Auto-Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ARMA	Auto-Regressive Moving Average
ARMAX	Auto-Regressive Moving Average model with eXogenous inputs
BESS	Battery Energy Storage System
BPL	Broadband over Power Line
CBA	Cost-Benefit Analysis
CBR	Case-Based Reasoning
CIS	Consumer Information System
CNFL	Compañía Nacional de Fuerza y Luz
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
CV	Coefficient of Variation
DCU	Data Concentrator Unit
DER	Distributed Energy Resource
DHW	Domestic Hot Water
DL	Deep Learning
DR	Demand Response

DRMS	Demand Response Management System
DSM	Demand-Side Management
DSL	Digital Subscriber Line
DSO	Distribution System Operator
DSSE	Distribution System State Estimation
DST	Daylight Saving Time
DT	Decision Tree
DTW	Dynamic Time Warping
EC	European Commission
EKF	Extended Kalman Filter
EKZ	Elektrizitätswerke des Kantons Zürich
ELU	Exponential Linear Unit
EPRI	Electric Power Research Institute
ES2050	Energy Strategy 2050
EU	European Union
EV	Electric Vehicle
EWZ	Elektrizitätswerk der Stadt Zürich
FAN	Field Area Network
FASE	Forecasting-Aided State Estimation
FHMM	Factorial Hidden Markov Model
GA	Genetic Algorithm
GAFAM	Google, Amazon, Facebook, Apple, and Microsoft
GBRT	Gradient Boosting Regression Tree
GDPR	General Data Protection Regulation
GIS	Geographic Information System
GMM	Gaussian Mixture Model
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
GUI	Graphical User Interface
HA	Hour-Ahead
HA	High-Activity

HAN	Home Area Network
HEMS	Home Energy Management System
HES	Head-End System
HMM	Hidden Markov Model
HVAC	Heating, Ventilation, and Air Conditioning
ID	Identification Number
IDE	Integrated Development Environment
IHD	In-Home Display
ILM	Intrusive Load Monitoring
IQR	Interquartile Range
IWB	Industrielle Werke Basel
KNN	K–Nearest Neighbor
LA	Low-Activity
LAV	Least Absolute Value
LDA	Linear Discriminant Analysis
LoRaWAN	Long Range Wide Area Network
LR	Linear Regression
LSTM	Long Short-Term Memory
LTE	Long-Term Evolution
LV	Low-Voltage
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MAR	Missing At Random
MA&ES	Moving Average and Exponential Smoothing
MCAR	Missing Completely At Random
MCM	Markov Chain Model
MDMS	Meter Data Management System
MFH	Multi-Family House
MILP	Mixed Integer Linear Programming
ML	Machine Learning
MLE	Maximum Likelihood Estimation

MNAR	Missing Not At Random
MPC	Model Predictive Control
MV	Medium-Voltage
NA	Not Available
NAN	Not A Number
NAN	Neighborhood Area Network
NDA	Non-Disclosure Agreement
NILM	Non-Intrusive Load Monitoring
NMAE	Normalized Mean Absolute Error
NN	Neural Network
NRMSE	Normalized Root Mean Square Error
NSHRP	Normalized Sharpness
NTL	Non-Technical Loss
OMS	Outage Management System
OPF	Optimal Power Flow
P2P	Peer-to-Peer
PCA	Principal Component Analysis
PF	Power Factor
PLC	Power Line Carrier
PLF	Probabilistic Load Forecasting
PMU	Phasor Measurement Unit
PV	Photovoltaic
QGBRT	Quantile Gradient Boosting Regression Tree
QLSTM	Quantile Long Short-Term Memory
QRNN	Quantile Recurrent Neural Network
REL	Reliability
RES	Renewable Energy Sources
RF	Radio Frequency
RL	Reinforcement Learning
RLC	Remote Load Control
RLU	Rectified Linear Unit
RMS	Root Mean Square

RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RT	Real-Time
R&D	Research and Development
SCCER	Swiss Competence Centres for Energy Research
SE	State Estimation
SHRP	Sharpness
SLA	Standard Load Allocation
SLP	Standard Load Profile
SM	Smart Meter
SoC	State-of-Charge
SVM	Support Vector Machine
SVR	Support Vector Regression
TCL	Thermostatically Controlled Load
TMCM	Traditional Markov Chain Model
ToU	Time of Use
TSO	Transmission System Operator
UKF	Unscented Kalman Filter
UMTS	Universal Mobile Telecommunications System
UTC	Coordinated Universal Time
WAN	Wide Area Network
WH	Water Heater
WLAN	Wireless Local Area Network
WLS	Weighted Least Square
WSN	Wireless Sensor Network
XML	Extensible Markup Language

LIST OF FIGURES

Figure 2.1 Representation of a smart grid with its typical components such as distributed generation, rooftop solar pannels, electric vehicles, and smart meters at the distribution and consumption side [19]. 14

Figure 2.2 Typical smart grid data volume growth (Terabytes) for a utility with one million customers [23]. 16

Figure 2.3 Overview of AMI communication networks [36]. . . . 20

Figure 2.4 Cost-Benefit Analysis for the deployment of smart electricity meters to at least 80% of all customers by 2020 (as of July 2018) [59]. 25

Figure 2.5 Overview of target period for the completion of a roll-out of smart electricity meters to at least 80% of all customers (as of 2018) [59]. 26

Figure 2.6 Grid stress test visualized in the Adaptricity platform [198]. 47

Figure 2.7 GridEye visualization platform of depsys [203]. . . . 48

Figure 3.1 CIGRE benchmark model of low-voltage distribution network [213]. 55

Figure 3.2 Schematic representation of distribution grid structures. 57

Figure 3.3 Daily roll-out statistics of smart residential meters in the City of Basel over 2015 and 2016, and number of additional and missing meter data with respect to the previous day. 58

Figure 3.4 Illustration of a digitized section from the distribution network of the City of Basel operated by IWB. . . . 60

Figure 3.5 Illustration of a LV section with corresponding loads from the sub-grid presented in Figure 3.4. 61

Figure 3.6 Illustration of a digitized section with corresponding loads from the distribution network of the City of San José operated by CNFL. 63

Figure 3.7 Typical weekly profile of the total consumption in the sub-grid presented in 3.6, split into residential, commercial, and industrial loads. 64

Figure 3.8 Representation of periods of measurement per advanced meter in the Costa Rican sub-metering study. 65
Figure 3.9 Load profiling of a single residential user over one day. 66
Figure 3.10 Load profiling of the aggregation of 70 residential users over one week. 66
Figure 3.11 Consumption share among residential appliances in the Costa Rican sub-metering study. 67
Figure 3.12 General pipeline for data preparation, transforming raw data into clean and tidy time series and metadata. 70
Figure 3.13 Example of different original structures of AMI measurement data sets in tabular form. 74
Figure 3.14 Illustration of measurement data set in B-format. . . 76
Figure 3.15 Histogram of the percentage of missing values in active power measurements of residential consumers in the City of Basel. 80
Figure 3.16 Histogram of the Pearson's correlation coefficient between the solar irradiance and the active power production of metered PV systems in the City of Basel. 81
Figure 3.17 Example of a time series with active power measurements where outliers are detected ($\alpha := 10$). 82
Figure 3.18 Relative difference between smart metered yearly energy consumption and billed energy as provided in metadata for residential consumers of a same neighborhood. 86
Figure 3.19 Imputation of a four-day gap in the load profile of an industrial consumer based on the mean, linear regression, replica from the following week, and ANN. 90
Figure 3.20 Illustration of conversion process from 10-minute to 15-minute resolution. 94
Figure 3.21 Weekly load profile of the aggregation of smart metered consumers in a LV grid and of the corresponding distribution transformer, together with their mismatch. 95
Figure 4.1 Iterative process of the k-means clustering algorithm to build three clusters out of the Fisher's Iris data set. 103
Figure 4.2 Graphical user interface for the selection of clustering features. 105
Figure 4.3 Clustering of normalized typical daily load profiles. . 108
Figure 4.4 Typical daily profiles of smart metered loads clustered according to their mean energy consumption. 110

Figure 4.5 Spatial representation of smart metered loads clustered according to their share of energy consumed at different periods of the day (early morning, morning, afternoon, evening, night). 113
Figure 4.6 Spatial representation of smart metered loads clustered according to their correlation with the temperature. 115
Figure 5.1 Active power load profiles over one day of a water heater, a tumble dryer, and a refrigerator at different temporal resolutions. 124
Figure 5.2 Box and whisker plots of the maximum value and of the coefficient of variation in active power measurements of multiple domestic appliances at different temporal resolutions. 125
Figure 5.3 Load profiling of the residential user presented in 3.9 and displayed at different temporal resolutions. . . . 127
Figure 5.4 Box and whisker plots of the maximum value and of the coefficient of variation in active power measurements of residential users at different temporal resolutions. Labels “with appliance” and “w/o appliance” indicate whether the indicators are computed with or without the presence of the specific appliances.128
Figure 5.5 Load profiling of the aggregation of 70 residential users displayed at different temporal resolutions. . . . 130
Figure 5.6 Smart meter load profiles at different spatial aggregation levels, scaled by the respective number of consumers.131
Figure 5.7 Sensitivity analysis of the impact of temporal resolution and spatial aggregation on the maximum power value and the coefficient of variation of load profiles. 134
Figure 6.1 Procedure for the synthesis of active and reactive power pseudo-measurements. 144
Figure 6.2 Scatter plots of reactive power with respect to active power for four different consumers from the Costa Rican smart meter data set. 156
Figure 6.3 Box and whisker plots of the power factor of 1’000 Costa Rican consumers categorized by the consumer size, period of the day, and active power level. Outliers are not shown for the sake of clarity. 158

Figure 6.4 Active power gap (i.e., yellow shape) over a typical summer and a winter day between the aggregated smart metered profiles and the transformer loading in the residential neighborhood presented in Figure 3.5. 164
Figure 6.5 Illustration of metered and synthetic active power profiles on two distinctive weekdays for a residential and a commercial consumer. 165
Figure 6.6 Load distribution of the original and synthetic profiles for a specific large consumer between 18:00 and 19:00. 166
Figure 6.7 Box and whisker plots of the mismatch in energy between metered and synthetic data, evaluated at different seasons and types of weekday. 167
Figure 6.8 Evolution over one year of the average power consumption of the original gap profile and of the aggregation of optimally selected synthetic profiles. 168
Figure 6.9 Illustration of reactive power synthesis over one day for a small and a large consumer, with and without scaling (i.e., spatial aggregate matching). For better visualization, active and reactive power values are normalized between 0 and 1. 170
Figure 6.10 Box and whisker plots of the root-mean-square error between metered and synthetic reactive power profiles, categorized by synthesis approach and consumer size. Outliers are not shown for the sake of clarity. 172
Figure 7.1 Average NRMSE over all bus injections in function of the share of yearly energy measured per bus by smart meters. 181
Figure 7.2 Component loading and voltage of the benchmark grid. 184
Figure 7.3 Structure of the sensitivity analysis according to the different dimensions and options under consideration. 186
Figure 7.4 Example of active power injection at a particular bus with respect to different SM placements, SM penetrations and types of pseudo-measurements. . . . 187
Figure 7.5 Error distribution of active power injection (before SE) at two specific buses between multiple scenarios with random SM placement and the benchmark model, considering different SM penetration levels and two approaches for active power pseudo-measurement synthesis. 188

Figure 7.6 Comparison of the median adjusted RMSE in active power injections between the strategic and 10 random SM placements for different levels of SM penetration, and different types of active power pseudo-measurement synthesis. 190
Figure 7.7 Adjusted RMSE of active power injections among the loaded buses with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements. . . . 195
Figure 7.8 Difference in 95^{th} percentile of active power injections among the loaded buses with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements. . . . 196
Figure 7.9 Adjusted RMSE of line loadings with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements. 198
Figure 7.10 RMSE of bus voltages with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements. . . . 199
Figure 8.1 Active power histogram of an example house (i.e., house 6 in Table 8.1) with and without refrigerator. 213
Figure 8.2 Histogram of the total active power consumption at various temporal resolutions in an example house (i.e., house 3 in Table 8.3) with and without a water heater. 214
Figure 8.3 Disaggregation process of a refrigerator load according to the low-activity-based approach based on 1-minute resolution data. For visualization purposes, power values are capped at 1500W. The illustrated section corresponds to the squared area in Figure 8.7. 218
Figure 8.4 Disaggregation process of a refrigerator load according to the jump-based approach based on 1-minute resolution data. For visualization purposes, power values are capped at 1500W. The illustrated section corresponds to the squared area in Figure 8.7. 218

Figure 8.5 Disaggregation process of a water heater load in an example house (i.e., house 1 in Table 8.3) based on 10-minute resolution active and reactive power measurements over one week. 221
Figure 8.6 Typical confusion matrix used as basis for classification metrics. 226
Figure 8.7 Disaggregation of the refrigerator load in an example house (i.e., house 2 in Table 8.1) over one day at different temporal resolutions according to both proposed approaches. For visualization purposes, power values are capped at 1000W. Blue-shaded areas are low-activity periods. 227
Figure 8.8 Performance evaluation of the refrigerator load disaggregation based on error metrics. 229
Figure 8.9 Performance evaluation of the refrigerator load disaggregation based on classification metrics. 230
Figure 8.10 Disaggregation of the water heater load in an example house (i.e., house 1 in Table 8.3) over three days at different temporal resolutions, with and without the help of reactive power (Q). 231
Figure 8.11 Performance evaluation of the water heater load disaggregation based on error metrics, accounting for the use of reactive power. 233
Figure 8.12 Performance evaluation of the water heater load disaggregation based on classification metrics, accounting for the use of reactive power. 233
Figure 8.13 Performance evaluation of the water heater load disaggregation on the Costa Rican sub-metering data set. 234
Figure 8.14 Average power consumption and share of disaggregated cold appliances and water heaters in the aggregation of the 70 household loads from the Costa Rican sub-metering data set, categorized by weekday and weekend. 236
Figure 9.1 Workflow followed in this work for the design of a forecasting algorithm. 243
Figure 9.2 Measurements and associated predictions based on the AMCM and SVR algorithms for two representative load profiles. 245

Figure 9.3 Principle of the swinging door algorithm with $\epsilon := 0.4$ and resulting extracted ramps. 252

Figure 9.4 Various statistical properties of the load profiles and respective predictions presented in Figure 9.2. The skewness and kurtosis values are scaled by the respective values calculated over the measurements. 253

Figure 9.5 Performance evaluation of different forecasting algorithms based on the MAPE, NRMSE, and two variants of the adjusted NRMSE. 257

Figure 9.6 Performance evaluation of different forecasting algorithms based on two variants of the ramp score. . . . 258

Figure 9.7 Performance evaluation of different forecasting algorithms based on the relative difference in the trend, seasonality, and autocorrelation with a lag of 1 day. . 258

Figure 9.8 Performance evaluation of different forecasting algorithms based on the relative difference in the skewness, kurtosis, and coefficient of variation. 259

Figure 9.9 Principle of the KNN algorithm used for quantile forecasting. 264

Figure 9.10 Weeks of the data set split into training set, validation set, and test set. 267

Figure 9.11 Quantile forecasts of active power consumption performed by the neural network at a specific bus with PV panels and EVs for scenario DS_3. Prediction intervals with 50% and 80% are visualized. 272

Figure 9.12 Quantile forecasts of active power flow at a specific line for scenario DS_3, accounting the car feature. Prediction intervals with 50% and 80% nominal confidence are visualized. 273

Figure 9.13 Quantile forecasts of voltage magnitude at a specific bus with PV panels and EVs for scenario DS_3, accounting for the car feature. Prediction intervals with 50% and 80% nominal confidence are visualized. . . . 274

Figure 9.14 Average coverage error of the quantile forecasts of active power flow for all lines and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered. [1] 275

Figure 9.15 Average interval score of the quantile forecasts of active power flow for all lines and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered. 276
Figure 9.16 Average coverage error of the quantile forecasts of voltage magnitude for all buses and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered. [2] 276
Figure 9.17 Average interval score of the quantile forecasts of voltage magnitude for all buses and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered. 277
Figure 9.18 Intermediate voltage point forecasts and final voltage realizations resulting from strategy S_{44} at a selected time step. 282

LIST OF TABLES

Table 3.1 Specifications of most common measurement data types in B-format. 72
Table 5.1 Average value of the indicators of the sensitivity analysis illustrated in Figure 5.7. Red and blue values refer to the sub-metering and smart meter data sets, respectively. 133
Table 6.1 Power factor of standard domestic appliances as reported in [287]. 157
Table 6.2 Average values of the power factors illustrated in Figure 6.3, categorized by intra-day period, consumer size, and active power level. 159
Table 7.1 Types of measurements in a distribution grid. Standard deviation values in italic are given in relative terms. 179
Table 7.2 Dimensions of the sensitivity analysis. 186
Table 8.1 Specifications of 14 refrigerators in the Costa Rican sub-metering data set. 224
Table 8.2 Statistics over 24 water heaters in the Costa Rican sub-metering data set. 224
Table 8.3 Specifications of 3 water heaters used in the evaluation. 225
Table 8.4 Number of successful disaggregations of refrigerator load according to the temporal resolution and the disaggregation approach. 229
Table 9.1 Performance evaluation of the forecasts of the two representative load profiles illustrated in Figure 9.2. . 250
Table 9.2 Overview of the different DER modified data sets used in the case study. 267
Table 9.3 Basic feature set for the probabilistic prediction of bus quantities. 268
Table 9.4 Basic feature set for the probabilistic prediction of line quantities. 268

Table 9.5 Optimization results for different voltage control strategies in current and potentially future price situations, considering the curtailed energy and the associated cost. 283

1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The last decade has experienced a rapid worldwide roll-out of new advanced sensor elements, especially so-called smart meters, in electrical distribution grids. This development is often triggered by national and sometimes supranational policies. For example, the Swiss Energy Strategy 2050 considers smart electricity meters as a key component of the future smart grid in Switzerland [1]. At the European level, the European Parliament expected the replacement of at least 80% of conventional electricity meters with smart electronic meters by 2020 wherever it was cost-effective [2]. Concretely, smart meters enable accurate high-resolution measurements on both the spatial scale (i.e., down to the end-consumer level) and the temporal scale (i.e., within the range of seconds to minutes) for parts of the distribution grid for which only spatially aggregated measurements (e.g., at the substation level) have been previously available. At first glance, the main motivation of power utilities to install smart electricity meters is the more efficient integration of billing data into the existing billing system by avoiding manual data gathering. Nevertheless, such large-scale digitalization, including end-consumers and producers, is additionally an excellent opportunity for a better operation and planning of distribution grids.

Traditionally, Distribution System Operators (DSOs) were used to monitor and operate their system on a medium-voltage level for aggregation of end-users, while the low-voltage grid was considered as a black box. Its infrastructure was usually over-dimensioned to cope with the worst-case scenarios according to the respective load. However, the recent and necessary transition towards a more sustainable energy system translates into the appearance of new electric loads like charging stations for electric vehicles and electric heat pumps, but also Distributed Energy Resources (DERs) such as rooftop Photovoltaic (PV) systems and battery storage units. On the one hand, these new elements connected at the low-voltage level put the system infrastructure under pressure, which implies new challenges for DSOs in terms of grid operation, notably for voltage control and congestion management. On the other hand, they are associated with certain flexibility that can be potentially leveraged for a more cost-efficient distribution grid operation and planning than the classical grid reinforcement measures. The distribution grid is not considered as a passive unobservable load anymore but tends to be an active component of the power system. This inevitably relies on a better comprehension of the low-voltage level. The previously unattainable degree of detail provided by the Advanced Metering Infrastructure (AMI) effectively allows for enhanced visibility and controllability if, and only if, power utilities possess suitable methods and tools for data processing, modeling, analysis, and visualization.

Consequently, the power system research community performs a multitude of data-based studies and suggests a large variety of data-based approaches to deal with the operation and planning challenges of active distribution grids [3–5]. Among others, various visualization, topology estimation, grid modeling, and state estimation techniques are suggested to increase the visibility and transparency in distribution grids [6–8]. Traditionally applied at the transmission level, load forecasting also becomes popular in distribution grids [9, 10]. In addition, load profiling, customer characterization, load disaggregation, and demand response schemes are proposed in the current literature as the basis for a more cost-effective grid operation [11–13]. Furthermore, privacy-preserving techniques and transactive energy systems which directly profit the end-users are emerging topics [14, 15].

Nevertheless, a large gap appears between state-of-the-art practices in the power industry and the various data-based applications proposed in the literature, where the overwhelming majority remains at the conceptual stage. On the one hand, the power industry is particularly conservative and tends to principally trust well-known and time-tested solutions (e.g., grid reinforcement). More importantly, system operators generally lack the expertise to properly make use of the gathered data and are certainly not aware of their full potential that is unfortunately largely under-exploited. On the other hand, data-based approaches proposed in the literature too often rely on unrealistic assumptions and can not be directly applied to real-world systems, especially:

- Full system observability, usually assuming a complete smart meter penetration and a perfect knowledge of the grid structure, is a prerequisite of many optimization and control schemes, which is totally impractical.

- AMIs data are always prone to inaccuracies, anomalies, and missing values which often prevent the direct application of proposed data-based approaches and inevitably impact their efficiency.

- A majority of models are inspired by the practices at the transmission level, especially in state estimation and forecasting. Their application to distribution grids, characterized by different properties and notably a much higher inherent uncertainty, is questionable.
- Load disaggregation techniques and demand response schemes often rely on measurements at the device level and/or of very high frequency which are usually not available in current distribution grids.
- The large majority of case studies are based on synthetic data and on simplistic test grids which do not reflect the much more complex reality of distribution grids.

This thesis aims to bridge some of the gaps which currently prevent power utilities and their customers from making use of the full potential of actual measurement data in distribution grids. For that purpose, the entire work is solely based on real-world data, and case studies are designed to be as realistic as possible. The characteristics, the potential, and the limitations of AMI data are comprehensively analyzed. In that respect, different data-based approaches are proposed to bypass the limitations of current metering systems while addressing some challenges faced by distribution system operators, energy providers, and aggregators. It must be noticed that the proposed work only focuses on applications relying on data with a sampling frequency of one minute or lower. Applications requiring higher resolution data are out of the scope of this thesis.

1.2 CONTRIBUTIONS

This work focuses on the use of data at the low-voltage level for grid operation and planning purposes. Specifically, the main contributions of this thesis are:

- Exclusive usage of real-world measurement data and grid models in the presented case studies, and discussion on the realism of data-based approaches and analyses proposed in the literature.
- Comprehensive description of the necessary preparation process for AMI data.
- Special focus on the visualization and interpretability of a large amount of AMI data.
- Analysis of the impact of temporal resolution and spatial aggregation on the characteristics of load profiles in distribution grids.

- Design of an approach for the synthesis of active power profiles with realistic properties at both the individual and the aggregate level.
- Discussion on reactive power pseudo-measurements and generation of synthetic reactive power profiles.
- Exhaustive sensitivity analysis of the principal dimensions influencing the data-based modeling of distribution grids down to the low-voltage level.
- Consideration of multiple properties (e.g., point-wise error, statistical properties) in the performance evaluation of state estimation and forecasting algorithms.
- Unsupervised detection and disaggregation of cold appliance and water heater loads based on standard smart meter data (i.e., at a temporal resolution between 1 and 30 minutes).
- Discussion on the adequacy of traditional deterministic load forecasting algorithms and standard evaluation metrics in low-voltage grids.
- Hour-ahead probabilistic forecasting of the state (i.e., net power consumption, power flow, voltage) of low-voltage grids.
- Design of a preventive voltage control scheme at the low-voltage level on the basis of quantile forecasts.

1.3 THESIS OUTLINE

The thesis is organized into three parts. The first part introduces the notion of data in distribution grids, elaborates on their potential usage, and gives insight into their characteristics and their interpretation in different contexts. The second part details how AMI data can be leveraged to model distribution grids, especially at the low-voltage level. The third part proposes different data-based applications for low-voltage grids. The content of each part is divided into the following chapters:

Part I

- **Chapter 2** discusses the digital transformation and its implications in electric distribution grids. The Advanced Metering Infrastructure (AMI) is presented, which relies on the installation of advanced metering devices such as the popular smart meter and the design of appropriate communication networks and data management systems. An overview of the roll-out of smart meters in Switzerland, in Europe, and in the World is given. Such digitalization gives rise to new applications but also brings new challenges which are reviewed in this chapter. In this context, an increasing number of start-up companies emerge to leverage in practice the large amount of data produced in distribution grids.

- **Chapter 3** first describes multiple real-world data sets leveraged for the purpose of this thesis, illustrating the diversity of data available in distribution grids. In the second part, the focus is given to the necessary preparation process of AMI data to ensure a certain quality before their use for further analysis.

- **Chapter 4** highlights the utility of unsupervised learning and of visualization to enhance the comprehension of large AMI data sets. In fact, the potential of such big data cannot be identified without appropriate data mining techniques for complexity reduction. A suitable representation of the extracted information provides power system engineers with valuable intuition in their decision-making process. As a case study, this chapter focuses on the k-means clustering algorithm and its application to smart meter data gathered over an entire city.

- **Chapter 5** analyzes the alteration of load profiles at the distribution grid level with respect to the temporal resolution and spatial aggregation. Although rarely properly considered in the literature, both temporal and spatial dimensions can highly influence the conclusions of data-based studies, especially at low aggregation levels.

Part II

- **Chapter 6** focuses on the creation of power profiles for end-consumers in low-voltage grids. The synthesis of pseudo-measurements at this level is indispensable to obtain an observable system and cope with the lack of direct measurements. A novel approach is proposed for the generation of active power profiles which are consistent at an aggregate level but also ensures realistic properties at the level of individual consumers. Rarely considered in the literature, the synthesis of reactive power is also thoroughly addressed. Besides, a methodology is developed to optimally allocate synthetic load profiles to actual non-metered consumers in a given distribution system.

- **Chapter 7** studies the influence of the AMI design and of the related modeling of pseudo-measurements on the outcome of distribution system state estimation, especially at the low-voltage level. A comprehensive sensitivity analysis is carried out that accounts for the type, the penetration level, and the placement of metering devices that compose state-of-the-art AMIs. The different techniques developed in Chapter 6 for the synthesis of power pseudo-measurements are also considered and compared against traditionally used approaches.

Part III

- **Chapter 8** suggests novel approaches for the detection and disaggregation of cold appliance and water heater loads. In contrast to non-intrusive load monitoring techniques generally suggested in the literature, the proposed approaches are unsupervised and only rely on commonly available smart meter measurement data with a temporal resolution between 1 and 30 minutes. These domestic appliances are characterized by a non-negligible flexibility potential such that an accurate estimation of their power demand at each instant shall notably contribute to more efficient demand response schemes.
- **Chapter 9** elaborates on the application of short-term forecasting in low-voltage grids. The high volatility and low predictability of the load at this level challenge deterministic forecasting algorithms and question the suitability of commonly used evaluation metrics. In contrast, probabilistic approaches allow for a certain quantification of the large uncertainty inherent to low-voltage systems. In this chapter, quantile forecasting is leveraged to estimate the near-future system state in a probabilistic way. Resulting quantile forecasts are subsequently integrated into the design of preventive voltage control schemes.

Finally, **Chapter 10** summarizes the key findings of this thesis and suggests directions for future work.

1.4 PUBLICATIONS

Part of the work presented in this thesis has been reported in the following publications:

[1] **Thierry Zufferey** and Gabriela Hug. “Impact of data availability and pseudo-measurement synthesis on distribution system state estimation”. In: *IET Smart Grid* 5.1 (2021), 29.

[2] **Thierry Zufferey**, Gabriela Hug, and Gustavo Valverde. “Unsupervised disaggregation of water heater load from smart meter data processing”. In: *Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower)*. IEEE. 2020, 1.

[3] **Thierry Zufferey**, Gabriela Hug, and Gustavo Valverde. “Disaggregation of cold appliance loads from smart meter data processing”. In: *PES Transmission & Distribution Conference and Exhibition-Latin America (T&D LA)*. IEEE. 2020, 1.

[4] **Thierry Zufferey**, Sandro Renggli, and Gabriela Hug. “Probabilistic state forecasting and optimal voltage control in distribution grids under uncertainty”. In: *Electric Power Systems Research* 188 (2020), 106562.

[5] **Thierry Zufferey**, Alice Lepouze, and Gabriela Hug. “Inadequacy of standard algorithms and metrics for short-term load forecasts in low-voltage grids”. In: *PowerTech*. IEEE. 2019, 1.

[6] **Thierry Zufferey**, Damiano Toffanin, Diren Toprak, Andreas Ulbig, and Gabriela Hug. “Generating stochastic residential load profiles from smart meter data for an optimal power matching at an aggregate level”. In: *Power Systems Computation Conference (PSCC)*. IEEE. 2018, 1.

[7] **Thierry Zufferey**, Andreas Ulbig, Stephan Koch, and Gabriela Hug. “Unsupervised learning methods for power system data analysis”. In: *Big Data Application in Power Systems*. Elsevier, 2018, 107.

The following papers have been published in the course of the PhD studies, but their content is not directly included in the thesis:

[8] Jochen Stiasny, **Thierry Zufferey**, Giacomo Pareschi, Damiano Toffanin, Gabriela Hug, and Konstantinos Boulouchos. “Sensitivity analysis of electric vehicle impact on low-voltage distribution grids”. In: *Electric Power Systems Research* 191 (2021), 106696.

[9] Jun-Xing Chin, **Thierry Zufferey**, Etta Shyti, and Gabriela Hug. “Load forecasting of privacy-aware consumers”. In: *PowerTech*. IEEE. 2019, 1.

[10] Gustavo Valverde, **Thierry Zufferey**, Stavros Karagiannopoulos, and Gabriela Hug. “Estimation of voltage sensitivities to power injections using smart meter data”. In: *International Energy Conference (ENERGYCON)*. IEEE. 2018, 1.

[11] **Thierry Zufferey**, Andreas Ulbig, Stephan Koch, and Gabriela Hug. “Forecasting of smart meter time series based on neural networks”. In: *International Workshop on Data Analytics for Renewable Energy Integration.* Springer. 2016, 10.

[12] Leandro Von Krannichfeldt, Yi Wang, **Thierry Zufferey**, and Gabriela Hug. “Online ensemble approach for probabilistic wind power forecasting”. In: *Transactions on Sustainable Energy* (under review).

[13] Yi Wang, Leandro Von Krannichfeldt, **Thierry Zufferey**, and Jean-François Toubeau. “Short-term nodal voltage forecasting for power distribution grids”. In: *Applied Energy* (under review).

In addition to the aforementioned publications, the content of this thesis is partially based on the following semester and master theses conducted in the Power Systems Laboratory of ETH Zurich:

Master theses:

[1] Benjamin Schaule. *Disaggregation of SmartMeter Measurement Data into Specific Load Components.* Dec. 2017.

[2] Carmen Exner. *Implementation of a Predictive Maintenance Engine for Power Cables.* Apr. 2018.

[3] Alice Lépouzé. *Design of an Automatic Forecasting Engine for Real-time State Estimation in Distribution Grids.* Apr. 2018.

[4] Haoyang Zhang. *Distribution Grid State Estimation based on Smart Meter Data.* Aug. 2018.

[5] Iraklis Katsolas. *Supply Curve Forecasting in the Italian Power Market.* Sept. 2018.

[6] Iason-Ioannis Chontzopoulos. *Very Short-term Probabilistic Demand Forecasting at High Aggregation Level for the Mitigation of Balancing Costs.* Nov. 2018.

[7] Filippo Ferrando. *Potential Impact of Electrical Vehicles on the Swiss Energy System.* Apr. 2019.

[8] Jochen Stiasny. *Sensitivity Analysis of EV Impact on Distribution Grids and Economic Assessment of Mitigation Strategies.* Apr. 2019.

[9] Sandro Renggli. *Robust State Estimation for the Operation of Distribution Grids with High Share of Distributed Energy Resources.* Aug. 2019.

[10] Tara Katamay-Smith. *Adaptive Model of EV Flexibility for Building and EV Load Optimization*. Sept. 2019.

[11] Leandro Von Krannichfeldt. *Online Ensemble Learning for Short-Term Load Forecasting*. Aug. 2020.

Semester theses:

[12] Alice Lépouzé. *Probabilistic Approach in Time-Series Simulation of Power Flows*. June 2017.

[13] David Rodriguez Flores. *Integration of a Dynamic Model for Stratified Electric Water Heaters into DPG.sim*. June 2017.

[14] Joël Dunant. *Investigation of Forecasting Techniques in Distribution Grids*. Jan. 2018.

[15] Stefan Agovski. *Synthesis of Reactive Power Profiles for Consumers in a Distribution Grid*. June 2018.

[16] Roman Engeler. *Long-term Energy Forecasting of Smart Metered Customers in a Distribution Grid*. July 2018.

[17] Sandro Renggli. *Distribution Grid State Estimation based on Forecasted Smart Meter Data*. Jan. 2019.

[18] Julie Rousseau. *Impact of Time and Spatial Aggregation of Smart Meter Data*. June 2020.

[19] Shipra Mohan. *Probabilistic Forecasting in Buildings and Privacy Concerns*. Jan. 2021.

[20] Timmy Frischknecht. *Coordination Strategy for Flexible Buildings*. Jan. 2021.

Part I

DATA IN DISTRIBUTION GRIDS

2

DIGITALIZATION OF DISTRIBUTION GRIDS

This chapter provides an overview of the digital transformation and its implications in electric distribution grids. This digital transformation is fostered and facilitated by the roll-out of advanced metering devices and the design of appropriate communication networks and data management system, which constitutes the advanced metering infrastructure. Smart meters are the most popular advanced metering devices and are being largely installed at the final electricity consumers. This large-scale roll-out gives rise to new applications but also brings new challenges which are largely discussed in recent literature. Furthermore, an increasing number of start-up companies emerge in the energy sector to leverage in practice the large amount of data produced in distribution grids.

2.1 INTRODUCTION

Digitalization or digital transformation can be defined as the use of digital technologies and the acquisition of digital skills to change a business model and provide new revenues and value-producing opportunities [16, 17]. The concept of digitalization has definitely revolutionized our daily life environment which tends to become always "smarter". The wide range of applications of a smartphone or in a smart home is the most apparent example. In that respect, observers state that we are currently experiencing a Fourth Industrial Revolution (also called Industry 4.0) [18]. Among others, rapid technology breakthroughs in the fields of Artificial Intelligence (AI), robotics, and Internet of Things (IoT) allow for a large-scale self-monitoring and automation of industrial processes with high gains in efficiency. This digital revolution transforms industrial sectors as diverse as manufacturing, medicine, biology, transportation, or even agriculture.

In this context, the energy sector is no exception. Reliability, efficiency, and adaptability requirements are transforming the gas, heat, and electric power systems into smart energy networks [20]. As illustrated in Figure 2.1, the notion of Smart Grid (SG) is particularly popular in the electric power sector.

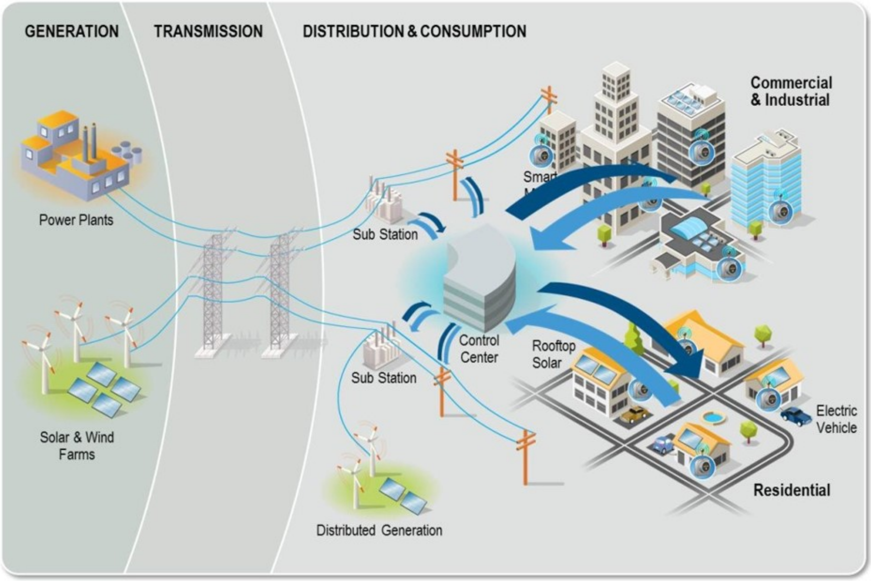
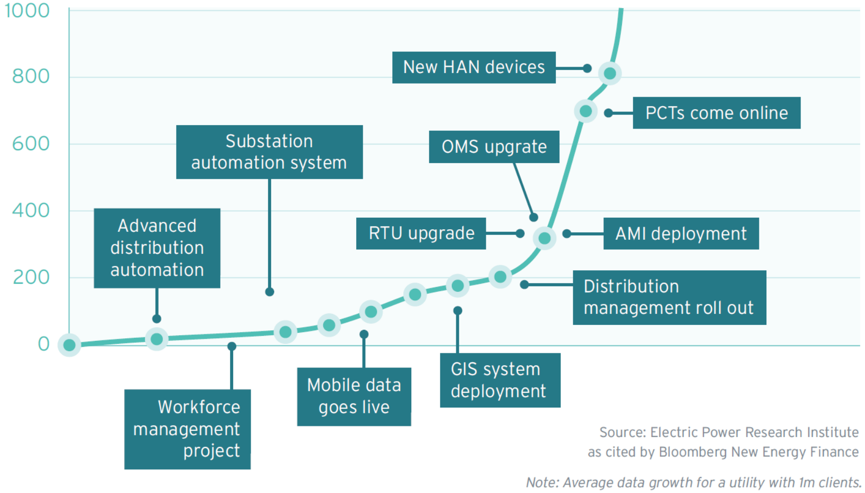


Figure 2.1: Representation of a smart grid with its typical components such as distributed generation, rooftop solar pannels, electric vehicles, and smart meters at the distribution and consumption side [19].

This notion mainly refers to the large-scale installation of high-resolution sensors with communication capabilities. Traditionally, only the large power plants, the transmission grid, and the substations have been monitored and controlled, notably to ensure voltage and frequency stability, whereas the distribution grid was seen as a black box. The power was known to flow from the large power plants to the end consumers in the distribution grid. In addition, the distribution grid infrastructure has usually been over-dimensioned to cope with any unexpectedly high load. There was no need to monitor the state of distribution grids. However, the recent aspirations for a more sustainable energy supply and a reduction of energy-related CO2 emissions lead to the decommissioning of large fossil fuel power stations. Distributed Energy Resources such as small wind turbines, solar Photovoltaic (PV) systems, and Battery Energy Storage Systems (BESSs) are being installed to compensate for the drop in energy supply from the transmission grid. Concurrently, consumers become prosumers, generate part of their power consumption, and acquire Electric Vehicles (EVs), electric heating systems, and electric water heaters, which additional loads that put the low-voltage grid under stress. The direction and magnitude of power flows are not anymore foreseeable. This introduces new paradigms in the operation and planning of distribution grids, where more efficiency, flexibility, and intelligence are generally required. The monitoring of distribution grids down to the end consumers becomes inevitable such that power utilities have initiated a large roll-out of high-resolution measurement devices since the beginning of the 21st century. This trend started in the commercial and industrial sectors and continues in the residential sector.

The smart metering system is an electronic system capable of measuring electricity fed into the grid, consumed from the grid, or flowing through the power lines [21]. There is a large variety of smart measurement devices, notably depending on the measured quantities and sampling frequency. The most common type of smart measurement devices are so-called Smart Meters (SMs) which are installed at end consumers. The capabilities of smart meters are also very diverse, but all devices basically record power values at a frequency spanning from one minute to one hour. In contrast to the conventional meters whose readings are manual and on-site, smart meter data are transmitted automatically to the responsible power utility, which facilitates the billing process. In addition, smart meters allow for better tracking of customers relocating. Moreover, the connection of customers to electricity can be automatically interrupted via smart meters in case of non-payment. While these are the primary interests of replacing traditional meters with smart meters, a smart metering system opens up a wide range of new business opportunities. On the one hand, end customers can obtain accurate feedback on their consumption and potentially adapt their consumption habits to save energy and lower their electricity bill, e.g., by shifting some of their consumption to lower price periods. On the other hand, system operators can leverage this source of information to better monitor the power flows in their grid and reduce the costs for grid operation and planning. Advanced smart meters are additionally equipped with control functionalities that allow for certain management of the customer's flexible energy resources. In addition to smart sensors and devices, a reliable and secure communication network and data management system are necessary to build the Advanced Metering Infrastructure (AMI) [4, 22].

"Big Data" are traditionally characterized by the so-called five V's, i.e., their volume, variety, velocity, veracity, and value. These characteristics also apply to data in distribution grids [24]. First, although it lies far below the massive data sets processed by Big Tech companies like the GAFAM (i.e.,



Source: Electric Power Research Institute as cited by Bloomberg New Energy Finance

Note: Average data growth for a utility with 1m clients.

FIGURE 2.2: Typical smart grid data volume growth (Terabytes) for a utility with one million customers [23].

Google, Amazon, Facebook, Apple, and Microsoft), the growing volume of data in a smart distribution grid is within the range of Terabytes, as illustrated by Figure 2.2 for a population of one million customers [23]. This requires a dedicated data management system and appropriate data analysis and machine learning tools, commonly based on distributed or cloud computing services.

Second, there is a great variety in the type and the source of the data. Single-phase or three-phase measurement time series of typical electric quantities such as the current, voltage, active and reactive power, and power factor form the majority of the data. These measurements mainly come from residential, commercial, and industrial consumers, PV and storage systems, transformers, and cable distribution cabinets. More and more individual electric devices, especially with some flexibility potential like water heaters, heat pumps, and EVs, are metered. Further quantities are thus being measured, e.g., battery State-of-Charge (SoC), battery charge and discharge rate, water temperature, room temperature. The metering devices generally differ among the sources, each having its own standards in terms of communication, sampling resolution, precision, and data formatting. In addition, the AMI integrates information about the distribution grid from a Geographic Information System (GIS), where the network topology, line and transformer parameters, status of switches, capacity of production and storage units, nominal power of consumers, as well as their geographic coordinates are stored. Exogenous data like weather measurements and forecasts, characteristics of customers, or energy market information, indispensable for a cost-efficient grid operation, add up to the variety of data.

Third, regarding their velocity, smart meter data can still not be defined as problematic. The databases of power utility are commonly updated once a day with the measurement data of the previous day. Generally, smart meters are characterized by an output frequency much lower than the measurement frequency which is typically in the kHz range. Even in cases where the updates are done in (quasi-)real-time, the output resolution of smart meters usually spans from one minute to one hour. This can be handled by power utilities and does not necessarily require the use of online processing algorithms. Nevertheless, Phasor Measurement Units (PMUs), which are standard devices in transmission grids, might also become popular in distribution grids [25, 26]. They provide ultra-high-resolution measurements of voltage phase angle and could potentially be used for diagnosis (e.g., unintentional island detection, state estimation, fault location, and oscillation detection) and control (e.g., protective relaying, Volt-Var optimization, and microgrid coordination) applications in distribution grids [27]. Being characterized by an output rate between 10 and 60 Hz, the roll-out of so-called micro-PMUs in low-voltage grids would be a real challenge for power utilities in terms of data synchronization and integration into the AMI. However, their currently high cost compared to the potential benefits hinders a large development in distribution grids.

Fourth, the veracity of data in distribution grids is a non-negligible aspect. All data sets are subject to noise, unintentional anomalies, unrealistic values, missing values, or missing timestamps due to a failure of the sensors, in the communication system, or in the database. A consistent data cleansing is primordial before further using the information for potential applications. In addition, energy theft, also referred to as Non-Technical Loss (NTL), costs billions of dollars annually on a national level to energy providers [28]. The most common methods are meter tampering (i.e., the meter is hacked to block or slow the accumulation of consumption statistics), meter bypassing, meter switching, tapping on low-voltage lines, etc. NTLs can be detected and limited by combining different AMI data sources [29].

Finally, the value of AMI data goes far beyond the sole simplification of the billing process and detection of energy theft. State Estimation (SE), load forecasting, customer characterization, Demand Response implementation, connection verification, or outage management are some of the applications of smart meter data analytics [4]. Besides these five main V's, the authors in [30] mention additional Big Data characteristics that apply to data in distribution grids, e.g., validity, variability, vulnerability, visualization. These different aspects will also be handled over the course of this thesis.

The rest of this chapter is structured as follows. Section 2.2 characterizes the advanced metering infrastructure and, more specifically, the design of smart devices, communication networks, and data management systems. Section 2.3 presents the roll-out of smart meters in Switzerland, in Europe, and in the World. Section 2.4 details the main applications and challenges of smart meter data. Section 2.5 lists a few Swiss companies which facilitate the digital transformation of distribution grids. Finally, Section 2.6 summarizes the main aspects and sets the basis for this thesis.

2.2 ADVANCED METERING INFRASTRUCTURE

In the case of a smart electric grid, Advanced Metering Infrastructure (AMI) is an integrated system of smart meters, communications networks, and data management systems that enables two-way communication between utilities and customers [31]. Smart meters and diverse advanced metering devices are used to monitor and control appliances at consumers' premises [20]. In addition, these metering devices send the collected information to the utility servers and receive operational commands from the operation center. This requires highly reliable and secure communication networks [22, 32]. Finally, the collected information is stored in a data management system to be processed and analyzed for billing purposes but also for more advanced applications. These three sub-systems of AMI are detailed in the following subsections.

2.2.1 *Smart Meters*

For decades, conventional electromechanical meters were the main type of device for measuring electricity flows. They are designed to count the number of revolutions of an electrically conductive metal disc that rotates at speed proportional to the power passing through the meter [33]. Measured data are usually displayed on an analog counter, and readings have to be manually recorded [20]. Electromechanical meters do not contain any further functionality. In order to facilitate the billing process and the tracking of customers relocating or switching to another energy provider, conventional meters have initially been replaced by Automated Meter Reading (AMR) devices. Mainly installed at industrial customers in the first instance, these devices allow for automatic and unidirectional communication of electricity consumption and production readings from the customer to the system operator and/or energy provider. The AMR data are typically sent once a month to ensure accurate billing.

Recently, the development of smart grids with increasing deployment of distributed energy resources and the need for supply and demand control at the customers' side inevitably requires the installation of advanced metering devices with bidirectional communication capability. Commonly called smart meters, they are electronic devices based on digital micro-technology. They rely on voltage and current sensors and do contain moving parts any more [34]. They record electricity quantities at fixed time intervals, typically between one minute and one hour, and transmit this information to the corresponding power utility in required time slots, typically once a day at night or sometimes even in near real-time. Modern smart electricity meters are open structures built in a modular way such that the main functions like metering and communication can be supplemented in a later stage by other modules with additional functions. In-Home Display (IHD) is an important module that allows customers to monitor their energy usage or production and obtain energy-saving feedback from power utilities [22]. Consumers can even be notified via IHD of an upcoming peak consumption event and be encouraged to temporarily reduce their own consumption. Nevertheless, IHDs contribute significantly to the total cost of smart meters and tend to be replaced by mobile applications. In addition, smart meters can be equipped with multiple communication modules. For example, the Home Area Network (HAN) transceiver enables to exchange information (e.g., receive sub-metering data and send control signals) with other electronic devices at home, whereas the GSM/GPRS module allows for communication with the outside [35]. Although smart electricity meters are normally supplied by the grid, an independent energy source (i.e., battery) can serve as a backup in case of a power outage.

2.2.2 *Communication Networks*

Figure 2.3 summarizes the different communication networks of which the AMI consists. First, in some customers' premises, Home Area Network (HAN) connects the main smart meter, which acts as an access point, with multiple

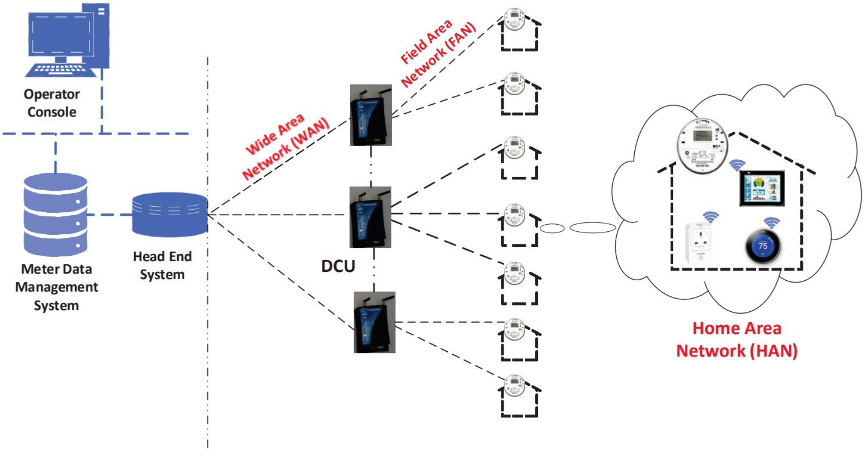


FIGURE 2.3: Overview of AMI communication networks [36].

smart devices such as energy storage (e.g., home battery) and generation (e.g., PV panel) units, EV charging installation, electric heat pump or water heater [22]. The data flow is instantaneous, the amount of data transferred is limited, and the network capacity should be extendable to new smart devices and a higher data rate. Given such requirements and considering the short distances between smart devices, low-power wireless technologies with low bandwidth (e.g., 10 to 100 Kbps per smart device) are currently the dominant solutions for HANs. In this case, the most common wireless technologies are WiFi, Bluetooth, ZigBee, Z-Wave, and 6LoWPAN [37–39]. Wired communication technologies such as Ethernet and HomePlug (i.e., technology that uses the existing home electrical wiring to communicate) are also being used. Based on HAN, a Home Energy Management System (HEMS) can monitor and manage electricity generation, storage, and consumption in a smart house [40].

Second, all smart meters in a specific area or neighborhood transmit their information to a Data Concentrator Unit (DCU). It is a gateway that facilitates bidirectional communication by relaying smart meter data to the data management system and passing on control commands from the system operator to smart meters [36]. DCUs are usually optimally placed in the LV grid, either in cable distribution cabinets[1], fixed on electricity poles, or

---

[1] Cable distribution cabinets are boxes installed at the LV level containing protection devices and where power lines are split.

even on a building, to reach all smart meters with a limited number of units and to meet the requirements in terms of delay and throughput [41]. The communication system between DCUs and smarts meters is called Field Area Network (FAN) or Neighborhood Area Network (NAN) and can have either a centralized or meshed topology [42–44]. NAN is typically based on small range coverage network technologies with a larger bandwidth than for HAN. Both wireless and wired technologies are suitable, such as Wireless Sensor Network (WSN) (i.e., smart meters act as a relay for other neighboring smart meters), Wireless Local Area Network (WLAN), Wireless Mesh Networks (WMN), Digital Subscriber Line (DSL), Power Line Carrier (PLC) communication, Broadband over Power Line (BPL), and even optic fiber cable. PLC has been the most implemented technology, especially in remote locations with low wireless coverage, since it directly uses the already existing grid [22]. Nevertheless, its low bandwidth (i.e., around 20 Kbps) is a disadvantage for a growing volume of data transferred such that broadband technologies are currently being preferred. In addition, the emergence of the 5G network infrastructure is seen as a very promising avenue for both HAN and NAN in order to ensure high reliability, high security, low power consumption, and high interoperability with an increasing number of devices [45].

Finally, Wide Area Network (WAN) allows DCUs, substations, power generation stations, and transformers to communicate with the utility's IT systems. In terms of technology, high data rate (i.e., up to 1 Gbps) and large coverage range (i.e., up to 100 km) are key requirements. For those reasons, WAN is mainly based on fiber optic, WiMAX, cellular communication (e.g., GPRS/UMTS/LTE), or more recently Long Range WAN (LoRaWAN) [44, 46]. The Head-End System (HES) acts as a hardware and software interface between the collection of AMI sensor data and the utility's IT systems.

2.2.3 *Meter Data Management System*

The Meter Data Management System (MDMS) is a system designed for long-term storage, management, and processing of smart meter data gathered by the HES [22]. Although its design largely varies among utilities based on their specific needs, the MDMS generally performs validation, cleansing, and analysis of smart metering data to support billing and decision-making processes. Moreover, MDMS interacts with multiple other IT systems such as the Consumer Information System (CIS), Outage Management System (OMS), Geographic Information System (GIS), and Demand Response Management System (DRMS) [47]. The CIS takes care of the relationship between utility and customers, and typically includes utility websites and the billing system whose efficiency is enhanced by the use of smart meter data. The OMS is responsible for the detection, location, and diagnosis of failures in distribution grids, as well as for the management and scheduling of restoration efforts. In this context, smart meters provide information that allows for faster and more precise fault location and diagnosis, and can alert customers regarding the restoration status [48]. Furthermore, smart meter data can be merged with traditional GIS data to comprehensively visualize the spatial distribution of consumption and production in the grid. Connectivity errors in GIS models can also be detected and corrected by leveraging smart meter data [49]. Finally, the DRMS largely profits from consumption data at the customer level in order to design specific demand response solutions [50].

2.3 ROLL-OUT OF SMART ELECTRICITY METERS

The roll-out of smart energy meters is considered an essential component of the energy transition and enables better monitoring and control of the energy systems. While the term "smart meter" can refer to electricity, gas, heat, and even water meters, this section only focuses on smart electricity meters, unless otherwise specified. The roll-out of smart electricity meters consists of replacing conventional meters with "smarter" electronic devices which primarily have the function of automatic remote reading [20]. Whereas some power utilities opted for an early installation of smart meters within the frame of pilot projects, the roll-out strategy is generally decided on the state level according to a Cost-Benefit Analysis (CBA). Its outcome largely differs across countries, which leads to very diverse smart meter penetration levels. In the following subsections, the roll-out of smart meters in Switzerland, Europe, and worldwide is detailed.

2.3.1 *Situation in Switzerland*

The national roll-out of smart meters in Switzerland is part of the Energy Strategy 2050 (ES2050) which has been designed by the Federal Council and Parliament, and approved by popular vote in 2017 [1]. ES2050 defines the scope of the Swiss energy roadmap until 2050. It includes measures for increasing energy efficiency, measures for the development of renewable energies, a ban on the construction of new nuclear power plants, and measures to speed up the modification and renovation of the electricity networks. In addition, Switzerland must no longer emit greenhouse gases that cannot be absorbed by natural and technical means (i.e., net zero emissions target) by 2050. Parliament also approved the development and operation of eight Swiss Competence Centres for Energy Research (SCCER), which guarantees additional funding and capacities in the field of applied energy research for Swiss universities [51].

Within the scope of ES2050, a new Federal Energy Act entered into force in January 2018, which specifies that 80% of the electricity meters must be smart meters by the end of 2027 [52]. The remaining 20% of meters can be used as long as they are functional. The decision on full smart meter roll-out arises out of a CBA carried out in 2015 on behalf of the Swiss Federal Office of Energy [53]. In brief, net costs (accounting for cost savings due to automatic meter readings) of 830 million Swiss francs must be expected for the setup and operation of the smart metering infrastructure between 2015 and 2035. These net costs are offset by electricity savings for end customers and savings in the business processes of electricity supply companies of 1'260 to 1'680 million Swiss francs. The nationwide roll-out of smart meters should therefore generate a quantifiable total net benefit of 430 to 850 million Swiss francs between 2015 and 2035. According to the Federal Energy Act, smart meter data with a resolution of 15 minutes or more must be handled in a pseudonymized way (i.e., personally identifiable information is replaced by pseudonyms) by DSOs, aside from its use for electricity billing. The use of pseudonymized smart meter data without explicit consent of the customers is nevertheless limited to measurement and control, implementation of tariff systems, secure and cost-efficient grid operation, setup of grid balancing, and grid planning. In addition, smart meter data can only be transmitted to third parties if they are pseudonymized or properly aggregated. If needed, the use of smart meter data with a higher resolution than 15 minutes is only allowed with the explicit consent of the customers. In any case, smart meter data must be deleted after 12 months unless they are still essential for billing or they are fully anonymized.

Most of the Swiss DSOs currently base their smart meter roll-out on the Federal Energy Act. For example, EWZ plans the replacement of 270'000 electricity meters from 2021 on in the City of Zurich for an estimated cost of 194.2 million Swiss francs [54]. However, a minority of DSOs have already initiated a large installation of smart meters before the approval of ES2050. Among others, this is the case of IWB that started the roll-out in the City of Basel in 2013 and already reached a 50% penetration in 2017 with almost 60'000 smart meters installed [55]. At this time, IWB was the DSO with the largest smart meter roll-out in all German-speaking countries. Similarly, EKZ also initiated their roll-out in 2013 in the canton of Zurich, and currently counts more than 150'000 smart meters in its network [56].

2.3.2 *Situation in the European Union*

Based on the directive 2009/72/EC of the European Parliament, at least 80% of conventional electricity meters should be replaced with smart electronic meters by 2020 whenever it is cost-effective [2]. For that purpose, EU Member States have been required to carry out a long-term CBA. As long as the CBA is negative, European states must perform a new CBA at least every four years. Once the result is positive, they must ensure the implementation of smart metering under EU energy market legislation within seven years [57]. In a first survey released in 2014 and benchmarking smart metering deployment in EU-27, the CBA of 13 states resulted in being positive, whereas the remaining states had still not conducted a CBA or have shown an inconclusive or negative CBA [58]. On this basis, Member States committed to rolling out close to 200 million smart electricity meters by 2020, which corresponds to 72% of European consumers. However, a second review released in early 2020 reveals a large discrepancy between planning and realisation [59].

Figure 2.4 illustrates the revised CBA results as of July 2018 for the deployment of smart electricity meters in EU-28. In addition, the target period expected by the Member States to achieve an 80% smart meter penetration is shown in Figure 2.5. In January 2018, 99 million (i.e., 34%) of all electricity customers in the EU-28 were equipped with a smart meter. Based on the observed rate of deployment in 2017, the authors in [59] estimate that about 24 million additional smart meters should be installed by 2020. This would correspond to a 43% penetration level, which is far below the initial expectations of 72%. The main reason for this gap is an insufficient regulatory framework at the level of the Member States, which does not fully ensure interoperability, data protection and security standards, or organizational effectiveness [60]. Other explanations are to be found in late approval of roll-out plans, political and/or financial instability, delays in starting the deployment, or technical and/or non-technical setbacks [61]. In fact, the situation largely differs among European states.

Among the states with positive CBA, Sweden was the first state to finish its deployment in 2009, followed by Italy in 2011, Finland in 2013, and Estonia in 2017. Currently, Sweden and Italy are even starting the roll-out of second-generation smart meters to replace the first generation which has

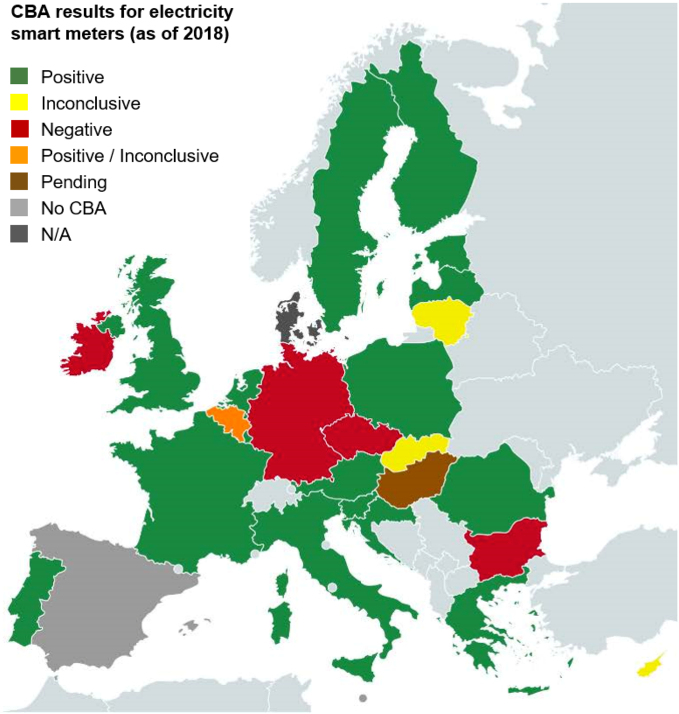


FIGURE 2.4: Cost-Benefit Analysis for the deployment of smart electricity meters to at least 80% of all customers by 2020 (as of July 2018) [59].

a technical and regulatory lifetime of about 15 years. In contrast to the first generation, this second wave must also comply with the requirements of the European Commission (EC) which recommends an output temporal resolution of 15 minutes [62, 63].

Although they did not conduct a CBA, Malta and Spain have directly chosen to go for full smart metering coverage, which has been reached in 2014 and 2018, respectively.

In France, all electricity meters are gradually being replaced by smart meters since 2015. Already 15.3 million (i.e., 44%) Linky meters have been installed by December 2018 and the total deployment of 35 million units is expected to last until 2021 [64].

Great Britain is far behind its schedule of full coverage by the end of 2020. Only 11.8 million smart electricity meters out of 28.6 million domestic and

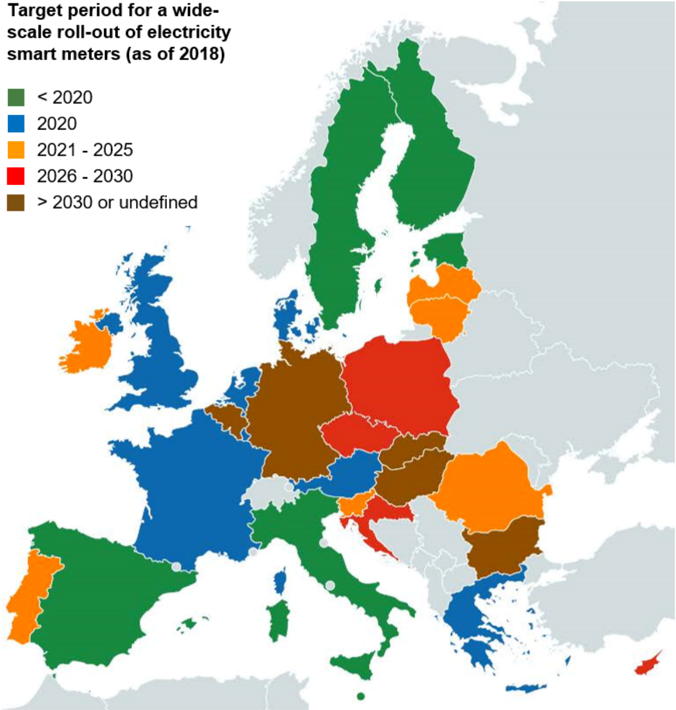

Target period for a wide-scale roll-out of electricity smart meters (as of 2018)
< 2020
2020
2021 - 2025
2026 - 2030
> 2030 or undefined

FIGURE 2.5: Overview of target period for the completion of a roll-out of smart electricity meters to at least 80% of all customers (as of 2018) [59].

non-domestic metering points (i.e., 41%) have been installed by June 2020 [65]. Among currently installed smart meters, 2 million devices must nevertheless be operated in traditional mode since suppliers are unable to operate them in smart mode or the meters cannot communicate via the Wide Area Network (WAN) at the point of reporting. Therefore, the government of Great Britain pushed back the deadline by four years to reach a penetration of at least 85% by 2024 [66]. It should be noted that energy suppliers must offer a smart meter to all UK customers, but customers can refuse their installation. In addition, the COVID-19 pandemic has a significant downwards impact on the number of new smart meter installations, which already threatens the latest roll-out target [67].

In Portugal, after being initially inconclusive, the second CBA conducted in 2015 turned out to be positive. A legal framework was only published in

2019, but the largest DSO already decided to deploy smart meters through large pilot projects. This led to 1.9 million (i.e., 31%) smart meters installed by the end of 2018 and aim to full coverage by 2025 [68].

Belgium commissioned region-specific CBAs and the outcome was inconclusive for Flanders, negative for Brussels, and had mixed results for Wallonia [61]. Therefore, there is no legally bounding target at the national level, but each region establishes its own roll-out strategy.

In Germany, the nationwide CBA carried out by Ernst & Young turned out to be negative. Therefore, the German authorities opted for a selective and step-wise roll-out based on the “Gesetz zur Digitalisierung der Energiewende” which came into force in August 2016 [69]. From 2017, all customers with consumption higher than 10’000 kWh per year must replace their traditional meter with a smart meter [70, 71]. From 2020, this annual consumption threshold is lowered to 6’000 kWh per year and producers with an installed capacity of at least 7 kWh must also be equipped with a smart meter. This corresponds to approximately 15% of all metering points. Private households with an electric vehicle, a heat pump, or PV panels might be concerned, but for most households (average electricity consumption of 3’500 kWh per year), smart meters are optional and only the installation of a modern measuring device (i.e., digital meter) is mandatory. The installation of modern and smart meters is staggered step by step until 2032 according to electricity consumption or generation capacity.

Finally, a few other countries, such as the Czech Republic or Bulgaria, simply do not plan a mandatory national roll-out after a negative CBA. Nevertheless, some DSOs still decided to progressively deploy smart meters in their network, especially when traditional meters reach the end of their technical lifetime [72, 73].

2.3.3 *Situation in the World*

Globally, IoT Analytics estimated an average smart meter penetration (electricity, water, and gas) of 14% in 2019 [74]. The consulting company also predicts a number of installed smart meters surpassing the 1 billion mark by 2021. Europe and North America are leading in terms of penetration rate, but Asia Pacific is clearly leading by overall volume. The rest of the world is at an early stage with low institutional support.

More precisely, North America was the first region in the world to move beyond traditional energy metering in the 1980s through the widespread introduction of Automated Meter Reading (AMR) devices. Nowadays, intelligent grids and smart meters are becoming an integral part of the development of smart cities. In 2019, the United States counted almost 95 million AMI smart electric meters, covering 60% of the 157 million electricity customers [75]. They are mainly installed at residential customers (88%), followed by commercial (11.4%) and industrial (0.5%) customers. Nevertheless, the smart meter adoption rate varies significantly among states. For example, Washington DC and Nevada are close to full penetration, whereas other states such as New Mexico or Utah barely reach 10% penetration. In Canada, over 82% of the meters are classified as smart meters in December 2018 [76]. This large-scale roll-out serves as the basis for a variety of smart grid projects going from distributed energy resource management to distribution grid monitoring and automation [77].

Asia-Pacific (mainly China, Japan, South Korea, India, Australia, and New Zealand) constitutes the world's largest and fastest-growing meter market with an estimated installed base of 618.8 million smart electricity meters in 2018 and annual demand in the range of 110–200 million units, where China accounts for about 70% of the volume [78]. The region is nevertheless highly fragmented in terms of the progress of smart metering deployments, and three general groups can be distinguished. China and New Zealand have more or less completed their first wave deployments of smart electricity meters. Second wave deployments are already underway in China and are soon to begin in New Zealand. South Korea and Japan are in the midst of their nationwide deployments and are scheduled to be fully deployed by 2020 and 2024, respectively. The third group consists of Australia and India which are in the early phases of smart meter deployments. The state of Victoria is, though, an exception with a large-scale roll-out already completed in 2013, i.e., 2.8 million smart meters covering 93% of households and small businesses [74]. In total, the penetration of smart meters in the Asia-Pacific region was 67% in 2018 and is expected to grow to 94% in 2024. This growth should primarily be driven by ambitious governmental targets in India to reach nationwide coverage within the next few years [78].

In Africa, Latin America, or the Middle East, the smart meter roll-out in a majority of countries is either still in a pilot stage or has not started yet [74]. In general, the main barrier to the adoption of smart meters in these regions is the lack of funding and government initiatives that have played a major role in the development of smart metering in other regions with larger penetration. In addition, inadequate infrastructure, based on obsolete technologies and often covering only urban areas, makes the deployment of smart meters still prohibitive for many utilities. Among African and Middle East countries, major implementations occur in Nigeria, South Africa, Egypt, United Arab Emirates, Saudi Arabia, and Qatar. In Latin America, Uruguay and Costa Rica are positioning to become the first countries with total coverage, which is expected by 2023 and 2024, respectively [79, 80]. Mexican utilities already achieved 10% penetration in 2019, with a longstanding goal of 30 million smart meters (i.e., 79%) by 2025 [81]. Chile is close to catching up with Mexico in terms of penetration and expects a complete roll-out by 2026. Finally, in Brazil, there is no nationwide roll-out plan and the deployment of smart meters is mainly restricted to pilot smart grid projects initiated by power distribution companies, especially in the region of São Paulo [82].

2.4 SMART METER DATA APPLICATIONS AND CHALLENGES

The availability of measurement data with high spatial (i.e., up to the end customers) and temporal (i.e., between one minute and one hour) resolution opens up a large set of opportunities for customers, electricity providers, and system operators. Recent advances in Machine Learning (ML) also help to deal with the massive amount of data gathered in distribution grids. Both supervised (e.g., linear/ridge regression, logistic regression, support vector machine, non-parametric regression, and decision tree) and unsupervised (e.g., principal components analysis, anomaly detection, and k-mean clustering) algorithms find applications in power distribution grids [83]. In addition, Artificial Neural Networks (ANNs), Deep Learning (DL), and Reinforcement Learning (RL) methods are becoming highly popular, especially applied to forecasting [10, 84].

In this context, smart data analytics becomes an important topic for all stakeholders in smart grids [4]. First, customers can better monitor and manage their energy consumption with the aim of reducing their electricity bills. In addition, prosumers can potentially trade their electricity production or share their DERs directly with other prosumers in the close neighborhood, which is known as transactive energy or Peer-to-Peer energy trading. Second, energy providers and aggregators can leverage smart meter data to design suitable price schemes, offer personal services, detect electricity theft, and better predict future demand. Third, DSOs benefit from high-resolution measurement data, which allows for higher transparency and visibility down to the LV grid. Distribution grid operation and planning can become more cost-efficient, notably with the design and implementation of appropriate Demand Response (DR) programs. Nevertheless, raw measurement data inevitably contain errors, noise, and missing values. As detailed in Chapter 3, consistent data preparation and cleaning is a necessary prerequisite before any further application. In addition, challenges in terms of data security and privacy protection arise, which needs to be addressed in order to guarantee the acceptance of customers for the use and analysis of personal information. Although most power utilities do not provide smart meter data to the scientific community for obvious privacy reasons, the release of a few open load data sets boosted the research on smart meter data analytics. A non-exhaustive list of open smart meter data sets is available in [4]. The following subsections explain in more detail some of the most relevant applications and challenges of measurement data in distribution grids, especially at the LV level, discussed in the literature.

2.4.1 *Detection of Non-Technical Losses*

Electricity theft is a serious issue in both developing and developed countries, which does not only induce revenue losses for energy providers but also jeopardize the distribution grid operation by underestimating the demand. For example, it has been estimated that electricity fraud is responsible for annual losses of up to 6 billion dollars in the United States and 4.5 billion dollars in India [85, 86]. Traditionally, Non-Technical Losses (NTLs) have mainly been caused by purely physical system manipulations such as the alteration of meter accuracy, bypassing of utility meters, or tapping of LV lines [87]. Although the introduction of smart electronic meters prevents physical meter tampering to some extent, the AMI becomes vulnerable to cyber-attacks. Electricity fraud in a smart grid can thus occur at all AMI levels, from the smart meter itself to the meter data management system as well as the communication systems.

Accounting for the large saving potential, the detection and location of energy theft became among the first concrete applications of smart meter data analytics, although it has been investigated long before the roll-out of smart meters via other means. An extensive literature on non-hardware electricity theft detection is available, especially based on smart meter data [88–90]. Fraud detection models can be categorized into four types: supervised classifiers, unsupervised approaches, state-based approaches, and game-theory-based approaches. Supervised classification algorithms such as Linear Regression (LR), Support Vector Machine (SVM), ANN, and Decision Tree (DT) ensemble are trained to distinguish normal from abnormal consumption patterns [29, 91–95], while addressing the challenge of imbalance data (i.e., abnormal consumers are usually much scarcer than normal consumers). Nevertheless, they require a large volume of labeled data, which is difficult to obtain. Alternatively, unsupervised anomaly detection, usually based on classical clustering algorithms or Gaussian Mixture Models (GMMs), enables to group consumers and isolate fraudulent users according to their behavior without the need for labeled data [96–98]. Fraudulent users with similar load profiles as normal consumers might however slip under the radar. Both supervised and unsupervised algorithms can be enhanced by the integration of exogenous data like the location or socio-economic information, although it may raise privacy concerns. Furthermore, state-based approaches leverage information from the grid such as voltage, current, and power flow measurements as input to a state estimation algorithm. Consumption data which are inconsistent with the estimated grid state are thus considered as abnormal [87, 99, 100]. Nevertheless, state estimation techniques require the network topology and a sufficiently large number of various and redundant measurements to obtain an observable system, which is rarely the case in low-voltage grids. Finally, game-theory-based approaches model the interaction of power utilities with benign customers and electricity thieves as a game, where the latter lead to different Nash equilibria than benign customers [101, 102]. Such a fraud detection scheme can be particularly efficient but is based on strong assumptions, notably regarding the type of fraud.

2.4.2 *Monitoring and Situational Awareness*

Traditionally, Distribution System Operators (DSOs) used to consider their power grid, especially the LV grid, as a black box since very few and only highly aggregated measurements (e.g., at substations) have been available. Distribution grids have often been over-dimensioned to cope with the worst-case scenarios. However, the emergence of new types of electrical devices (i.e., EVs, electric water heaters, and heat pumps) and DERs creates new challenges for DSOs, notably in terms of voltage control and management of line and transformer loading. Hence, properly modeling and monitoring the grid down to the lowest voltage level has become crucial in order to take appropriate and cost-effective operational and planning decisions. In this sense, high-resolution measurements coming from smart meters and diverse advanced metering devices installed at end-consumers, PV systems, BESSs, cable distribution cabinets, and distribution transformers help to achieve better visibility and transparency [5].

First of all, proper data visualization is a primordial step in the creation of valuable information and the extraction of knowledge from raw measurement data. In line with the big data challenges mentioned in the introduction, dedicated tools should be able to gather, integrate, and display the large and diverse amount of measurement data in an interface easily interpretable by energy data analysts and system operators, ideally close to real-time. For that purpose, different offline or cloud-based frameworks and dashboards are proposed in literature [6, 103–105]. They rely on big data mining algorithms such as simple statistical metrics (e.g., mean energy consumption, peak load value, standard deviation, percentiles) but also ML techniques (e.g., clustering, dimension reduction, anomaly detection, Deep Learning). Chapter 4 further elaborates on the challenges to visualize a large volume of smart meter data and details how clustering techniques help to decrease the data complexity and create useful information. Moreover, distribution grid monitoring tools are designed to provide system operators with a comprehensive spatial overview of their system state based on measurement data and supported by GIS data if available. Since distribution grids, including the LV level, are relatively large and complex systems, the information can be seen on multiple aggregation levels, going from the load and voltage profiles at substations down to the single customers. Hence, these software tools are interactive such that the user can zoom in or out but also click and hover on substations, lines, transformers, distribution cabinets, and even single customers to obtain specific and more detailed information. In addition to displaying recorded and GIS data, such tools might also detect and correct missing or erroneous values in the measurement data but also create synthetic load profiles to complete the set of measurements. More information on data preparation and realistic load profile synthesis is given in Chapters 3 and 6. On this basis, power flow simulation can be conducted to estimate non-measured quantities. In case of voltage band violation or component overloading, warnings and alerts are also typically triggered, often under the form of a traffic light system indicating the severity of the problem. Besides their monitoring function, advanced monitoring tools allow for the creation of future scenarios (e.g., increase in PV or EV penetration), spot potential contingency locations, and suggest preventive measures. Section 2.5 details more concretely different supportive monitoring tools for DSOs designed by Swiss start-up companies active in the design of smart grid solutions.

Furthermore, the existence of GIS information and of digital LV grid models is still not given for many DSOs. Hence, topology learning algorithms for distribution grids have been developed based on available measurement data [106–108]. The Swiss start-up company depsys has implemented a model-free topology learning algorithm based on their own LV sensors to support

DSOs in their digitalization, as detailed in Section 2.5.2. Current trends nevertheless indicate that an increasing number of DSOs start to digitalize their MV and LV networks into GIS. But digital grid models are still prone to errors such as connectivity errors or inaccurate line parameters, which can be verified and rectified with the help of smart meter data, sometimes combined with additional power flow measurements [109–111]. Assuming that the network topology is known, distribution grids can be operated in different configurations and DSOs are usually not aware of the status of switches, which can also be identified by data-based approaches [7, 112, 113]. Topology learning, verification, correction, and identification usually rely on graph theory and Maximum Likelihood Estimation (MLE). The algorithms typically leverage the fact that voltage measurements provide insight into the grid structure and lines' properties in the close neighborhood. Note, however, that a substantial portion of the approaches proposed in the literature assume full penetration of reliable smart meters, which is far to be the case in practice.

Finally, widely used in transmission grids, State Estimation (SE) is gaining in popularity in distribution grids. Accounting for the particularities of MV and LV grids (e.g., low x/r ratio, unbalanced conditions, radial configuration), numerous techniques are proposed in the literature to determine the most probable system state (i.e., bus voltage magnitudes and angles, bus active and reactive power injections, and active and reactive power flows) based on available measurements. Weighted Least Square (WLS), Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF), Least Absolute Value (LAV), Forecasting-Aided State Estimation (FASE), and ML techniques are among the most cited approaches [8, 114]. In contrast to the transmission level, measurements are generally insufficient in distribution grids to obtain an observable system which is a prerequisite for SE. Although smart meters provide a large part of the necessary measurements, these are prone to errors and full coverage is rarely achieved. Therefore, optimal sensor placement, bad data detection, and pseudo-measurements synthesis are key topics in Distribution System State Estimation (DSSE). In addition, distribution grid topology is often complex with a great number of nodes but also uncertain, as mentioned previously. Because of the observability problem and the large uncertainties in measurement data and grid models, DSSE is still rarely implemented in real systems. Chapters 6 and 7 specifically cover the synthesis of realistic load profiles and the practical challenges in terms of input data when implementing DSSE.

2.4.3 *Energy Forecasting*

Energy forecasting is among the most popular field of ML applied to power systems and mainly consists of load, solar power generation, wind power generation, and electricity price forecasting [10]. An impressively large number of scientific publications are available, going from review papers to the adaptation of existing supervised algorithms or the design of new energy-focused predictors. Load forecasting is the dominant category with about half of the publications in energy forecasting [115, 116]. Solar power forecasting is also highly popular since the early 2010s and the increasing share of energy produced by PV systems [117, 118]. While mainly relevant in transmission grids, wind power forecasting has been extensively studied, and the methodology can be adapted to load forecasting at the building level, accounting for the highly variable nature of both types of data [119, 120]. Although not directly linked to measurement data in distribution grids, electricity price forecasting is fundamental in electricity markets and is expected to have an increasing influence on the load in distribution grids, especially in the frame of DR programs [121, 122].

The forecasting methodology highly depends on the type of data and on the application. The type of data to predict particularly impacts the set of input features. For example, solar power forecasting is obviously mainly influenced by solar irradiance, whereas load forecasting profits from historical data and temporal information. In addition, smooth profiles are typically easier to predict and require less computational effort and less sophisticated algorithms than highly volatile profiles. In load forecasting, the volatility and predictability of the data are highly correlated with the aggregation level [123, 124]. Furthermore, the forecasting horizon and the class of predictors are determined by the application. Short-term predictions (i.e., day-ahead, intraday, hour-ahead, intra-hour) are traditionally used for operation purposes, whereas medium-term and long-term predictions (i.e., month-ahead, year-ahead, and decade-ahead) serve planning objectives [125, 126]. Prediction algorithms can be further split into deterministic and probabilistic approaches. Deterministic or point forecasting algorithms predict a single value per time step and aim to minimize the error with the real value. In contrast, probabilistic algorithms provide a probability distribution, a set of quantiles or prediction intervals per time step which inform on the uncertainty associated with the forecast [127, 128]. Research on probabilistic energy forecasting is principally pulled forward by wind power forecasting [119]. Based on the principle of “wisdom of the crowd”, forecast combination and ensemble forecasting generally allow for more accurate predictions than individual forecasting models, which is beneficial to both deterministic and probabilistic approaches [129–131]. Moreover, the hierarchical structure of power systems is totally appropriate for hierarchical time series forecasting which ensures coherent outcomes between predictions at different hierarchy levels [132]. Especially, the sum of all individual load forecasts in a grid must be equal to the aggregate load forecast at the feeder. Unfortunately, this knowledge is still barely leveraged in load forecasting [133].

Research in load forecasting has mainly focused on (highly) aggregate data at the transmission or substation level. However, the recent accessibility to smart meter data currently boosts the studies on forecasting in distribution grids down to the load of individual end-consumers. In their review of algorithms and models used for deterministic building energy consumption, the authors in [9] cite Artificial Neural Network (ANN), Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Machine (SVM), Case-Based Reasoning (CBR), Fuzzy techniques, Grey theory, Moving Average and Exponential Smoothing (MA&ES), K–Nearest Neighbor (KNN) as well hybrid methods. A growing number of publications also focus on residential load forecasting. Nevertheless, a majority of publications follow the same forecasting methodology for smart meter data as for aggregate data, although the properties vary significantly. On the one hand, the behavior of individual consumers is particularly hard to predict such that deterministic load forecasting does not provide reasonable information [134]. On the other hand, scoring functions based on the point-wise error and traditionally used for smooth and aggregate data are not suitable for volatile smart meter data [135, 136]. Probabilistic Load Forecasting (PLF) algorithms and the corresponding evaluation metrics are more appropriate to account for the large uncertainty at the building or household level. The related literature is still very limited and mainly focuses on the adaptation of deterministic models to quantile forecasting. For example, the authors in [137] suggest the use of Quantile Gradient Boosting Regression Tree (QGBRT), Quantile Recurrent Neural Network (QRNN), or Quantile Long Short-Term Memory (QLSTM). Among possible applications, short-term residential PLF can be used in the implementation of Demand Response programs and in transactive energy, whereas long-term residential PLF can be leverage in the sizing of home batteries and of the LV grid infrastructure. Chapter 9 is specifically dedicated to the challenges of deterministic and probabilistic forecasting based on smart meter data. Very rarely considered in the literature, voltage forecasting at the LV level is also taken into account in this thesis and leveraged for preventive voltage control.

2.4.4 *Demand-Side Management*

Demand-Side Management (DSM) refers to the general modification of the load. While energy efficiency measures (e.g., better building insulation, more efficient domestic appliances) target medium- to long-term reduction of energy consumption, Demand Response (DR) programs encourage the end-consumers to make short-term reductions in their power demand. The main objective is to provide ancillary services, and notably adjust the demand to intermittent energy production (e.g., solar and wind infeed), reduce forecast errors in a balancing group, maintain voltage within operational limits, and decrease consumption peaks which are harmful to the distribution system infrastructure [13]. DR can be achieved by financial incentives (i.e., discouraging energy consumption at certain hours of the day with higher prices than average) or by direct control of certain devices whose consumption pattern can be shifted without significant impact on the user. Thermostatically Controlled Loads (TCLs) such as the refrigerator, space heating, Water Heater (WH), or Heating, Ventilation, and Air Conditioning (HVAC) system are particularly appropriate for providing ancillary services thanks to their thermal inertia [138]. The flexibility of wet appliances like washing machines, tumble dryers, and dishwashers can also be exploited in a time window predefined by the user, although their rather sporadic usage limits the DR potential [139]. Home batteries also serve DR purposes by acting as a buffer between the actual consumption and the demand seen by the grid, which can additionally be combined with privacy-preserving functions [14]. Due to its high power and energy consumption but also flexibility potential, EV charging is recently often considered as a key element in DR schemes [140].

In this context, smart meter data help to identify good DSM candidates, enhance the implementation of DR programs, design efficient price schemes, and estimate the flexibility potential of consumers. The company O-Power is often cited as one of the first power utilities to leverage smart meter data for that purpose. It has implemented behavioral DR in order to reduce the overall energy consumption during the highest usage hours on peak energy days [141]. More precisely, O-Power utilizes the competition spirit of its customers by comparing their consumption behavior and encouraging them to outperform similarly situated customers in terms of consumption reduction during critical time periods.

The information revealed by smart meter data allows aggregators and system operators to personalize their energy efficiency measures and DR programs in order to maximize the propensity of customers to react as desired. More precisely, load profiling techniques allow for consumer classification according to their consumption pattern. Clustering approaches like k-means, fuzzy clustering, hierarchical clustering, and Self-Organization Mapping (SOM) are typically used for that purpose [11]. Feeding the clustering algorithm with complete load time series is however not very efficient. By reducing the set of features, dimensionality reduction techniques such as PCA contribute to better clustering [142]. Manual features extraction also enables the obtention of clusters according to the needs. For example, consumers can be grouped by k-means clustering according to the correlation between consumption and outside temperature, which gives insight into their probability to possess electric temperature-sensitive devices, as explained in Section 4.4.2. Moreover, load profiles contain socio-economic information that can be revealed with the help of customer characterization techniques. The authors in [143, 144] have shown that classifiers (e.g., KNN, Linear Discriminant Analysis (LDA), SVM, AdaBoost, and Convolutional Neural Network (CNN)) and regressors (e.g., multiple linear models) can be trained on smart meter data to identify a large variety of personal characteristics. For instance, the type and approximate age of the house, the floor area, the number of bedrooms and appliances, the type of cooking facility, or even the approximate age and the social class of the chief income earner can be identified with relatively high accuracy. Of course, the utility of knowing some of the cited characteristics is questionable, and privacy concerns are inevitably raised. Note that load profiling and customer characterization techniques are not only limited to DSM but can also serve other purposes such as improving load forecasting or bad data detection.

Furthermore, better insight into the composition of the load is necessary, and notably regarding the presence of low-efficiency devices that should be replaced or of flexible appliances that could be leveraged for DR purposes. Intrusive Load Monitoring (ILM) refers to the direct load recording of specific electric appliances in a building or household, also referred to as sub-metering [145]. Although the sensing and communication technology is available, ILM is relatively costly, requires additional installation and maintenance, and raises data privacy concerns, which prevents a wide-scale implementation of ILM. Alternatively, Non-Intrusive Load Monitoring

(NILM) uses the load profile at the building or household level to disaggregate the load of individual appliances. Literature on NILM is extensive and generally focuses on the detection of appliance-specific signatures. The NILM algorithms can be unsupervised, mainly based on Hidden Markov Models (HMMs), or supervised, such as Deep Learning approaches, which requires the availability of labeled data for training [12]. Nevertheless, state-of-the-art NILM techniques are computationally expensive and are currently not applicable on a wide scale since they rely on much higher temporal resolution data (i.e., a sampling rate of at least 1 Hz) than the resolution offered by currently rolled out smart meters (i.e., between one minute and one hour). Research on load disaggregation based on standard smart meter data is still scarce and principally targets the temperature-sensitive part of the load (e.g., coming from HVAC systems [146] or heat pumps [147]). Further discussion is given in Chapter 8, where fully unsupervised disaggregation algorithms are proposed.

Besides, the implementation of DR programs requires a good estimation of the flexibility potential on the demand side, more precisely regarding the duration and amplitude of possible load shifting, decrease or increase. If available, fine-grained data at the appliance level can be used to estimate their flexibility potential [148, 149]. Though, accounting for the lack of sub-metering data, most flexibility estimation studies make use of survey information which can only give a broad estimate at an aggregate level. Based on a regression model between outside temperature and smart meter data at the building level, the authors in [150] can still approximate the activity of AC systems and quantify the expected power reduction in response to a broadcast set-point change. Nevertheless, the actual flexibility of consumers can generally hardly be assessed based on smart meter data due to too high uncertainty regarding the precise functioning of flexible devices and the reaction of consumers. In addition, many studies try to model the price-elasticity of consumers [151, 152] or to design optimal Time-of-Use (ToU) or critical peak price schemes based on measurement data [153, 154]. However, it has been shown that consumers do not make consistently rational decisions as expected by those models [155, 156]. Therefore, a few pilot projects have been carried out to evaluate the actual response of consumers to DR programs. For example, the flexibility of wet appliances (where users set a deadline for the end of the program), EVs (where users set an expected departure time), and Domestic Hot Water (DHW) buffers (where comforts settings define the DR controllability range) has been quantified over 186 households in the LINEAR project [139]. It comes out that the flexibility potential is highly asymmetric (i.e., the power increase surpasses the power decrease potential) and significantly varies over the day. Similarly, a pilot study carried out in Norway over 40 households has investigated the potential of Remote Load Control (RLC) for WHs and the response to price signals in the use of wet appliances via a token indicating peak hours [50].

In practice, direct load control and Time of Use (ToU) tariffs are relatively common in the industrial sector [157]. However, despite the large flexibility potential in the commercial and residential sectors [158, 159], DR programs for those consumers are generally limited to simple ToU schemes, typically fixed peak and off-peak electricity prices to reduce the difference between the peaks and troughs of the demand profile. Nearly all advanced load management solutions proposed in the literature for commercial and residential consumers remain at the concept stage and are rarely put into practice for multiple reasons. The authors in [160–162] review the different barriers to an efficient implementation of DR programs.

First, the non-optimal behavior of consumers is considered a fundamental limitation. Although financial motivation is a key aspect, electricity is a relatively cheap good, and savings on customers' electricity bills may be insufficient to cover investments in equipment and to compensate for the inconvenience of participating in the program [163]. A study in [164] also found that the majority of participants in a DR program continue with their consumption habits and everyday routines despite regular energy feedback on their IHD. This highlights the need for proper information on the purpose and benefits of DR programs to ensure higher customer engagement. Quantifying the value and cost savings of DR is however associated with high uncertainty due to the difficulty to compute the baseline demand (i.e., the load in the absence of DR event), especially for residential consumers [165].

In addition, there is a clear lack of an appropriate market and regulatory framework. Nowadays, market structures are still too restrictive and usually require DR to be planned several hours ahead in addition to set high performance standards, which is often not realisable [166]. Current regulations also set restrictions on locational and temporal price differentiation and prevent transparent transmission of price signals to the final consumers such that they cannot perceive the true value of DR [167].

Finally, technological barriers are still limiting practical implementation. On the one hand, DR is based on a reliable AMI which is still in the roll-out phase in many distribution networks, as detailed in Section 2.2. Sensing, communication, computation, and control technologies have proven their efficiency in small-scale pilot projects. Nevertheless, there is still a substantial gap for wide-scale implementation of DR, notably in terms of experience for larger systems, penetration level of bidirectional smart meters, communication and computational capability, and harmonization of standards and protocols among the large diversity of systems and devices. On the other hand, proficiency skills in data science applied to power systems are required to properly understand and efficiently make use of the data gathered in distribution grids. Too many DR schemes proposed in the literature are still based on assumptions which are not in phase with reality. For example, some studies assume certain flexibility at an aggregate level without considering the practical feasibility and the implication for single consumers, or rely on perfect knowledge of the system, the demand, and the flexibility potential.

2.4.5 *Transactive Energy Systems*

Electricity markets traditionally perform resource allocation and pricing based on the conventional top-down approach of power system management, where customers and prosumers are passive receivers. But recently, so-called transactive energy systems have been proposed to properly integrate the increasing DER production [15]. These systems rely on a consumer-centric and bottom-up perspective by giving the opportunity to consumers and local producers to freely choose whom they want to exchange electricity with and what is the price they are willing to pay and offer, respectively. The design of such electricity markets is above all defined by the degree of decentralization and their topology. It goes from full Peer-to-Peer (P2P) markets with purely bilateral exchanges to community-based markets where multiple prosumers can collaborate in a microgrid. In this context, measurement data at the prosumer level are essential to account for the amount of power to be traded. In some cases, smart meters even serve as an interface to the trading platform through their in-home display.

Multiple local energy trading projects have already been successfully implemented [168]. For example, Piclo is an online energy marketplace in the UK that uses meter data, generator pricing, and consumer preference information to match electricity demand and supply every 30 minutes [169]. Similarly, Vandebron is an online platform in the Netherlands that allows consumers to directly buy electricity from independent producers such as farmers who own wind turbines in their fields [170]. In addition, blockchain technology is getting popular for facilitating the exchange of electricity without the mediation of a utility company or a financial institution [171]. In practice, LO3

Energy launched in 2017 in Brooklin the first microgrid energy market based on blockchain technology [172]. In Switzerland, the first blockchain-based energy market has been designed by the start-up company Exnaton whose activities are detailed in Section 2.5.3.

2.4.6 *Customer Concerns and Data Privacy*

Although smart meters offer a large set of opportunities for DSOs, aggregators, and energy retailers, they are also a source of worry among the customers. A serious frond against the installation of these devices is observed around the world, notably led by the organization “Stop Smart Meters” [173]. The organization provides support to customers who want to opt out or refuse the installation of smart meters and carries out multiple actions such as media outreach and street protest to discredit smart meters. It consists of many local associations, mainly active in the United States, but also in Canada, Mexico, Australia, New Zealand, Japan, and in European countries such as the United Kingdom, Portugal, Austria, or Norway. In France, in contrast to a majority of countries, customers do not have the right to refuse the installation of the Linky smart meter. This led to multiple legal proceedings against Enedis (i.e., the main French DSO) but also against recalcitrant customers [174].

Campaigns against smart meters are diverse and usually refer to health, reading accuracy, safety, and privacy concerns. First, inhabitants can be exposed to electromagnetic fields, more precisely Radio Frequency (RF) radiation, in case the smart meter is equipped with wireless communication. While this is definitely problematic for people suffering from electromagnetic hypersensitivity, like any exposure to electric devices, the RF radiation of smart meters is seen as a general health issue by their detractors. According to the American Cancer Society, it is very unlikely that living in a house with a smart meter increases the risk of cancer [175]. Nevertheless, there is no comprehensive long-term study regarding further health problems due to smart meters. It should still be noted that the amount of RF radiation from a smart meter is much lower than the radiation from a cell phone. Second, a few cases of fire hazard supposedly caused by defective or poorly fitted energy meters have been reported, although the operators ensure that meters cannot explode or ignite spontaneously [176, 177]. Third, some consumers have complained about higher electricity bills after the replacement of their traditional meter. Indeed, the authors [178] have shown that electromagnetic interference (e.g., caused by light dimmers or active infeed converters as used in PV systems) with the Rogowski coil current sensor of smart meters can lead to inaccurate, sometimes substantially higher, readings. Recent devices are nevertheless immune to such interference. Fourth, consumers do not directly see the benefits of smart meters over the traditional meters and fear that the installation cost may be indirectly passed on to them via the electricity bill. Finally, smart meters raise data protection and privacy concerns which are detailed in the following.

Due to the fine granularity of smart meter data, highly private information can be revealed about the behavior and habits of the electricity consumer with simple off-the-shell techniques, even without *a priori* knowledge of its activities or prior training based on sub-metering data. In that respect, Section 2.4.4 mentions the potential of customer characterization and NILM techniques to detect socio-economic features and the types of equipment being used, respectively. Occupancy profiles, lifestyle patterns, potential illnesses, and religious practices can be inferred from fine-grained consumption data. For example, the authors in [179] point out that hourly household consumption data can potentially reveal if a worker is at home during sick leave and if he got a good night's sleep by analyzing power activities during the day and at night, respectively. Minute- to second-resolution data can inform whether he ate a cold or hot breakfast, left late for work, or left his child home alone. The authors in [180] even show that audiovisual content can be detected based on household power consumption sampled every two seconds. Although some pieces of information are innocuous, other pieces are critical if third parties other than power utilities, like the employer, insurance companies, or even criminals, could have access to it. In theory, many components of the AMI (e.g., HAN, NAN, WAN, DCU, and MDMS) are vulnerable to cyber-attacks such as data interception and modification [181]. In addition, experiments have shown that vulnerabilities in specific smart metering infrastructures can be exploited by threat actors [182, 183].

According to a survey published by PwC in the US, customers attach high importance to cybersecurity and privacy but consider energy companies among the less trusted utilities in those fields [184]. This highlights the need for energy companies to put cybersecurity at the forefront of their business strategy and to build trust through action by implementing robust data governance. Countermeasures and protective actions against cyber attacks such as pseudonymization and encryption of meter data can already provide a certain level of cybersecurity [185]. It is also fundamental that regulatory standards in the roll-out of smart meters explicitly consider the benefits for consumers, and not only for metering companies, energy providers, and system operators. As experimented in the Netherlands at the beginning of the 2010s, the absence of a framework or set of safeguards in the standardization process to protect public interests can trigger a strong public resistance against smart metering [186]. In the Dutch case, smart meter roll-out was initially mandatory with legal consequences in case of refusal, which was considered as an infringement of the right to privacy as protected in the European Convention on Human Rights.

Recently, more stringent privacy protection laws, which also apply to smart meter data, have been adopted. This is the case of the European Union's General Data Protection Regulation (GDPR) which classifies the data gathered by smart meters as personal data that belongs to the customer [187]. Among others, GDPR stipulates that data subjects must be informed about processing operations, can restrict them under certain conditions, and have the right to ask for their data to be erased. In addition, privacy settings in the development of services must be set by default to a high level. However, there is still no clear guidelines regarding the way power companies should deal with customers to improve their acceptance of smart meters. For that purpose, the authors in [188] promote the role of ethics as an important driver in smart grid programs. In a so-called ethical smart grid, customers should be treated as collaborators, enabling a fruitful and long-lasting relationship between utilities and customers. Requirements in terms of ethics should rely on the concepts of parsimony and equity, and the risks of privacy breaches might be compensated financially.

Furthermore, the level of detail provided by smart meters must be the result of a trade-off between the benefits for power utilities and the privacy implications for consumers. First, data granularity plays a role in the possible detection of activities and appliance use. A higher temporal resolution allows for more detailed information, which increases the performance of NILM algorithms [189]. In the European Union, the granularity of smart meter data is therefore limited to 15 minutes, which automatically prevents the detection of most domestic appliances. Similarly, the aggregation level at which data is transmitted and processed also determines the degree of privacy protection. For example, the consumption data of a block of houses can be sent in an aggregated way to the utilities in order to preserve the privacy of the single households. Moreover, smart meter data is typically sent once a day to the main utility servers, usually at night for technical reasons. Such time-shifting of the data transmission prevents real-time monitoring of the consumers' activities. In general, it is clear that the possible applications and the associated performance depend on the level of detail in the data. For example, higher data resolution allows for more accurate DR program marketing, and the availability of real-time data improves short-time forecasting.

Unfortunately, privacy implications of the algorithms based on smart meter data and proposed in the literature are too often neglected. It is however possible to develop privacy-preserving approaches for some applications. For example, the authors in [190] have designed a k-means clustering algorithm that neither discloses an individual's private information nor leaks the community's characteristic data. In addition, the presence of a home battery can be leveraged to hide the household's consumption pattern seen by the meter using various charge and discharge schemes [191]. The authors in [14] have even proposed an energy management method that utilizes energy storage and local generation to simultaneously reduce energy costs and protect privacy through the minimization of information leakage. Nevertheless, other applications are hardly compatible with proper privacy protection. The authors in [192] have shown the limitations of privacy-preserving approaches in collaborative forecasting. In this case, cooperation between different data owners may increase the forecast performance, notably for wind and solar power, but at the cost of a loss of privacy despite protective measures. In the different studies, approaches, and applications presented in this thesis, privacy implications are explicitly considered. Specifically, Chapter 5 provides more information about the influence of temporal resolution and spatial aggregation on the level of detail visible in AMI data. In addition, the implications of temporal resolution for load disaggregation are discussed more concretely in Chapter 8.

2.5 SWISS START-UP COMPANIES AND SMART GRID SOLUTIONS

A large range of new business models has arisen with the digitalization trend observed in distribution grids, and especially the large-scale roll-out of smart meters. Traditional power companies usually lack the expertise to properly make use of the gathered data, and a portion of the customers wish to actively participate in the energy transition. Hence, many start-up companies are emerging in the energy sector. Only in Switzerland, about 250 energy start-up companies have already been created over the past ten years in the areas of transport technologies and services, energy production, energy-efficient technologies, storage and grid services, building technologies, energy management and use, and energy supply services [193]. They are highly active in R&D projects to address practical concerns that distribution grid operators and planners, energy retailers, and aggregators may face, but also directly focus on the interests of consumers and prosumers. For example, Hive Power provides a smart grid analytics platform (e.g., forecasting of energy demand, designing of new tariff schemes, and optimal management of energy communities and aggregated flexible loads) to energy suppliers and grid operators [194]. Zaphiro has developed a monitoring and automation solution to help power utilities integrate more clean energy technologies in the electricity grids while maintaining a high quality of service for their customers and optimizing system costs [195]. Bitblumens distributes solar power devices in areas without a power grid and connects them to the blockchain [196]. Clemap provides energy meters that analyze through load disaggregation techniques the energy consumption of buildings and propose energy efficiency measures [197]. In the following subsections, three start-up companies located in Switzerland and providing services to DSOs and/or prosumers are described in more detail. This gives a real-world insight into current applications in the field of smart grid solutions on the basis of measurement data in distribution grids.

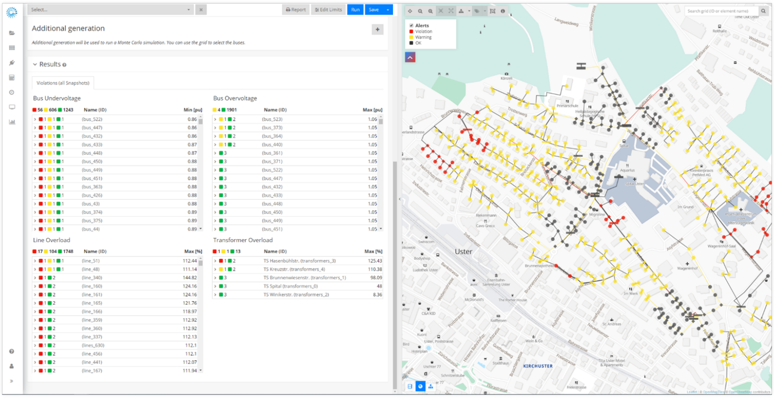
2.5.1 *Adaptricity*

Adaptricity is a spin-off company of ETH Zurich founded in 2014 and located in Zurich, describing itself as a driver of SmartGrid innovation in the German-speaking world [198]. The company took part in multiple start-up competitions and won several national and international awards, such as the CIRED Startup Award and the Asian Utility Award. Furthermore, it recently got awarded by the Watt d'Or 2021, a quality seal for energy excellence given by the Swiss Federal Office of Energy. The majority of the 60+ customers of Adaptricity are system operators, principally in Switzerland (e.g., IWB in the City of Basel, and EKZ and EWZ in the Canton and City of Zurich, respectively), but also in Germany (Bayernwerk in the region of Bayern), Austria (Netz Oberösterreich in the north of the country), and even Hong Kong (CLP Group) and Australia (AusNet Services in Victoria).

The business model of Adaptricity is built around grid analytics tools combining traditional grid planning practices with data-driven algorithms to leverage measurement data available in the power distribution grid. The Adaptricity platform is a simulation engine with a wide variety of grid calculation functions, and notably power flow time series simulations [199]. It basically consists of three products, namely Adaptricity.Plan, Adaptricity.Sim, and

Adaptricity.Mon, which are built upon the same cloud-based platform, and Adaptricity.Connect. Adaptricity.Plan focuses on the day-to-day operations of grid planning, such as power flow or short circuit calculations, grid reinforcement, or connection requests. Grid models can be easily imported using a wide variety of different data connectors. Adaptricity.Sim delivers a detailed analysis of the distribution grid based on time-series simulations. Different dashboards support the visualization of simulation results under the form of time series and statistics. The simulator is not limited to power-consuming appliances and traditional power generators but also integrates prosumer models, dynamic-pricing behaviors, SmartGrid applications, power-to-heat, etc., for an in-depth insight into the distribution grid of the future. Adaptricity.Mon provides comprehensive grid monitoring through the use of smart meters and substation measurement equipment. The visualization and evaluation of measurement data help the detection of operational violations and can spot negative trends in the grid. Since all measurement data are linked to a grid model, power flow calculations can be made for each time step, giving accurate information about the grid's operational state (i.e., voltages and line loadings) completely automatically. Hence, grid operators can visualize, simulate, and analyze their electricity grids in near real-time. This means more efficient grid operations, which in turn leads to better integration of Renewable Energy Sources (RESs) and fewer grid losses. In addition, Adaptricity.Connect allows end-customers to evaluate their connection requests themselves, saving DSOs and their customers significant time and effort. The preliminary calculation of RES hosting capacity per node determines which connection node is suitable for a new RES installation. The expensive options can be ruled out from the beginning, and the final connection request can be implemented cost-effectively. In this case, time-series-based grid simulations provide valuable insights for PV connection request assessment which cannot be obtained by purely static power flow simulations [200].

Furthermore, Adaptricity offers consulting services and tailor-made solutions, usually based on the Adaptricity platform. Currently, DSOs are making decisions partially blind since they have very little visibility into LV grids. Hence, load flow calculations based on smart meter data give insight into the grid's status-quo such that existing grid bottlenecks can be identified. In addition, the Adaptricity engineers can define plausible future grid scenarios for electric mobility and RES expansion. On this basis, detailed identification of grid bottlenecks for all plausible future grid scenarios can be performed. For example, the result of an LV grid stress test is illustrated in Figure 2.6. In addition, Adaptricity makes use of Monte Carlo simulation to assess the



Additional generation

Additional generation will be used to run a Monte Carlo simulation. You can use the grid to select the buses.

Results

Violations (all Snapshots)

Bus Undervoltage

🟥 56 🟨 606 🟩 1243	Name (ID)	Min [pu]
🟥 1 🟨 1 🟩 1	(bus_522)	0.86
🟥 1 🟨 1 🟩 1	(bus_447)	0.86
🟥 1 🟨 1 🟩 1	(bus_432)	0.86
🟥 1 🟨 1 🟩 1	(bus_433)	0.87
🟥 1 🟨 1 🟩 1	(bus_448)	0.87
🟥 1 🟨 1 🟩 1	(bus_450)	0.88
🟥 1 🟨 1 🟩 1	(bus_449)	0.88
🟥 1 🟨 1 🟩 1	(bus_451)	0.88
🟥 1 🟨 1 🟩 1	(bus_363)	0.88
🟥 1 🟨 1 🟩 1	(bus_426)	0.88
🟥 1 🟨 1 🟩 1	(bus_43)	0.88
🟥 1 🟨 1 🟩 1	(bus_374)	0.89
🟥 1 🟨 1 🟩 1	(bus_375)	0.89
🟥 1 🟨 1 🟩 1	(bus_44)	0.89

Bus Overvoltage

🟨 4 🟩 1901	Name (ID)	Max [pu]
🟨 1 🟩 2	(bus_523)	1.06
🟨 1 🟩 2	(bus_373)	1.05
🟨 1 🟩 2	(bus_364)	1.05
🟨 1 🟩 2	(bus_440)	1.05
🟩 3	(bus_361)	1.05
🟩 3	(bus_371)	1.05
🟩 3	(bus_522)	1.05
🟩 3	(bus_447)	1.05
🟩 3	(bus_432)	1.05
🟩 3	(bus_433)	1.05
🟩 3	(bus_448)	1.05
🟩 3	(bus_450)	1.05
🟩 3	(bus_449)	1.05
🟩 3	(bus_451)	1.05

Line Overload

🟥 17 🟨 104 🟩 1748	Name (ID)	Max [%]
🟥 1 🟨 1 🟩 1	(line_51)	112.44
🟥 1 🟨 1 🟩 1	(line_48)	111.14
🟥 1 🟩 2	(line_340)	144.82
🟥 1 🟩 2	(line_160)	124.16
🟥 1 🟩 2	(line_161)	124.16
🟥 1 🟩 2	(line_165)	121.76
🟥 1 🟩 2	(line_166)	118.97
🟥 1 🟩 2	(line_359)	112.92
🟥 1 🟩 2	(line_360)	112.92
🟥 1 🟩 2	(line_337)	112.13
🟥 1 🟩 2	(lines_630)	112.1
🟥 1 🟩 2	(line_456)	112.1
🟥 1 🟩 2	(line_441)	112.07
🟥 1 🟩 2	(line_167)	111.94

Transformer Overload

🟥 1 🟨 1 🟩 13	Name (ID)	Max [%]
🟥 1 🟩 2	TS Hasenbühlstr. (transformers_3)	125.43
🟨 1 🟩 2	TS Kreuzstr. (transformers_4)	110.38
🟩 3	TS Brunnenwiesenstr. (transformers_1)	98.09
🟩 3	TS Spital (transformers_0)	48
🟩 3	TS Winikerstr. (transformers_2)	8.36

FIGURE 2.6: Grid stress test visualized in the Adaptricity platform [198].

uncertainty inherent in future scenarios, such as the impact of electric vehicles on the grid infrastructure [201]. Finally, investment costs (e.g., grid reinforcement measures) can be estimated based on expected violations of standards. Grid Reinforcement measures are sometimes inevitable, but Adaptricity has shown that smart local control schemes can already substantially reduce the number of undervoltages in highly loaded LV grids [202]. Furthermore, Adaptricity develops interactive dashboards as stand-alone solutions or as web applications according to customer needs and use-cases in grid operation, planning, and asset management.

2.5.2 *Depsys*

Depsys is a start-up company founded in 2012 and located in Puidoux, in the French-speaking part of Switzerland [203]. Among other prizes, the cleantech company got awarded by the Solar Impulse Foundation and also received the Watt D'Or. GridEye is the core platform of depsys, already used by more than 40 grid operators mainly located in Switzerland, like Romande Energie and IWB. Based on both hardware and software components, GridEye allows depsys to produce and leverage high-precision, real-time data within the same platform, which decreases the risk of data leakage and errors. On the hardware side, field devices such as micro-PMUs are installed at key locations, usually at the LV side of distribution transformers and at cable distribution cabinets, for data acquisition and control. Measurements of three-phase electrical quantities are processed by the distributed intelligence on every device such

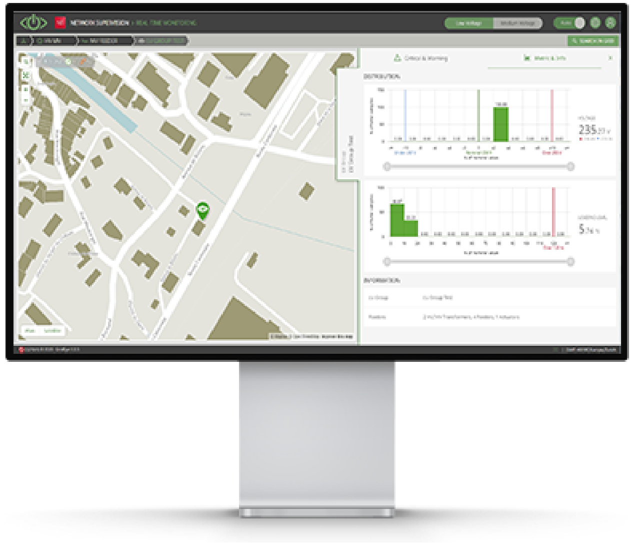


FIGURE 2.7: GridEye visualization platform of depsys [203].

that only useful data for the system operators and/or end-customers are communicated and stored. At the heart of GridEye, the management system is in charge of the IoT communication and data center. Finally, a variety of software modules give insight into the operation of the grid. The grid monitoring module provides real-time visibility into the condition of the grid and can send warning messages ahead of critical events. The power quality module handles supply compliance, troubleshooting, and root cause analysis. The fault management module triggers real-time alarms in case of interruption of supply and can rapidly determine the fault location. Temporal profiles (e.g., transformer daily loading, voltage and current variations) and statistics (e.g., transformer power, current, and voltage average daily or weekly dynamics, distribution of transformer loading, distribution of currents and voltages throughout the grid) are also visualized. In addition to the monitoring applications of GridEye, depsys offers further services to DSOs such as transformer aging analysis, imbalance and losses analysis, phase discovery, and optimal energy storage sizing. Figure 2.7 gives insight into the GridEye visualization platform.

In contrast to the Adaptricity platform, GridEye does not work with a GIS model of the grid, and network parameters are not known a priori. In addition, it does not necessarily rely on a large roll-out of smart meters, but principally on the high-resolution field devices developed by depsys. In order to compensate for the absence of a digital grid model, as is often the case at the LV level, depsys patented a method for estimating the topology of an electric power network using only well-placed high-resolution metering devices [204]. The method is based on the estimation of mutual current sensitivity coefficients and on an algorithm to obtain the network incidence matrix from the estimated sensitivity coefficients. In addition, depsys patented a model-less technique for determining mutual voltage sensitivity coefficients between a plurality of measuring nodes in an electric power network [205]. This approach has been successfully validated in laboratory for unbalanced LV grids [206]. The sensitivity coefficient estimation technique has then been leveraged in a research project to automatically determine grid hosting capacity based on measurement data. Without knowing the grid configuration, it provides insight into the impact in a distribution grid of any new installation (e.g., PV system, EV charger, storage system) and of the addition or replacement of grid components (e.g., cables, transformers). In another research project, GridEye measurements at the LV side of distribution transformers are used to create a digital twin of the MV side of the transformers [207]. Based on these digital twins, depsys further developed a new state estimation framework for real-time MV grid monitoring without the need for pseudo-measurements and the deployment of extensive and expensive grid measurement devices [208].

2.5.3 *Exnaton*

Exnaton is a very recent start-up company founded in Zurich in summer 2020 after winning a starting capital from Venture Kick [209]. The three co-founders have developed a software tool that creates local Peer-to-Peer (P2P) energy communities for trading renewable energy in the neighborhood. Based on the analysis of smart meter data, the software is intended to serve both energy providers and consumers. On the one hand, an application visualizes the electricity consumption and/or production of households, PV system owners, and small businesses in real-time. In addition, the software allows them to share electricity with each other. For example, a household can buy electricity from its neighbor’s PV systems. Prices for electricity are based on the availability of locally produced energy, which may save costs by targeted consumption behavior. On the other hand, the software platform calculates transactions among electricity customers and provides energy providers with easy-to-access billing data. New data-based revenue streams are also offered to energy providers, which ranges from self-consumption optimization up to the analysis of optimal sizing for PV systems.

The software platform builds on the experience and know-how gained in the research project Quartierstrom, where the future Exnaton team members could design and test the first local energy market of Switzerland in the frame of their PhD studies at ETH Zurich and University of St. Gallen [210]. In this pilot project, 35 households and two commercial entities from Walenstadt, a small Swiss village in the canton of St. Gallen, have joined a local energy community that allows the exchange and remuneration of electricity between consumers, prosumers, and the local electric grid provider without intermediaries. The community members pay a reduced tariff for grid usage if the electricity produced by a prosumer is sold to another community member located on the same voltage or grid level. Although the Swiss legislation does not currently support such novel location-grid pricing schemes, the pilot project has proven that such pricing structure can incentivize local balancing, i.e., locally produced energy can be consumed locally whenever possible to avoid costs from higher grid levels. Blockchain technology is used for logging the produced and consumed units of energy within the community [171]. Hence, both prosumers and consumers can indicate a price at which they are willing to sell or buy locally produced solar energy without third-party intermediaries. In addition, the pilot project has shown that the active involvement of households in the P2P energy market impacts their behavior and contributes to the energy transition [211]. Indeed, the community fosters sustainable practices (e.g., self-consumption or load-shifting) and the local aspects of the electricity exchange seem to drive user engagement, which might therefore facilitate the future diffusion of DERs.

2.6 CONCLUSION

To sum up, the electricity sector does not escape from the digitalization wave that revolutionizes many industry areas as diverse as medicine, biology, or manufacturing. Although the power industry is known to be particularly conservative, the wide-scale roll-out of smart meters currently transforms the distribution grid down to the low-voltage level which is not anymore seen as a passive black box. Beyond the sole installation of smart meters at customers' premises, further advanced measurement devices might be installed in distribution grids, e.g., at specific flexible devices, distribution transformers, and cable distribution cabinets. These various measurement devices, together with suitable communication networks and data management and storage systems, build the advanced metering infrastructure. Such infrastructure allows for a large variety of new potential applications which can serve all stakeholders in distribution grids, from electricity providers and system operators to end consumers. Besides automated meter reading which substantially simplifies billing processes, detection of non-technical losses is among the first practical applications seen by power utilities in regions where electricity theft is a serious issue. In addition, a widespread installation of such sensors highly contributes to increased visibility into distribution grids down to the LV level. Among others, measurement data are leveraged for visualization, monitoring, topology estimation, grid modeling, situational awareness, and state estimation purposes. Furthermore, access to smart meter data has boosted the research on energy forecasting, and especially load forecasting. These data can also be used for load profiling, customer characterization, load disaggregation, and estimation of flexibility potential, which enables the design of more cost-efficient demand response programs. Finally, transactive energy systems are getting popular in recent years, which profits directly the end consumers and prosumers who become active traders of their energy on the basis of accurate measurement data.

Nevertheless, while the detection of non-technical losses, grid monitoring, load flow simulations, and transactive energy systems are some applications that are successfully put into practice and commercialized, above all by start-up companies, a majority of applications are only at the conceptual stage. Indeed, there is often a large gap between the assumptions taken in data-based approaches proposed in the literature and the reality of measurement data in distribution grids. First of all, the general assumption of full smart meter penetration is currently unrealistic. In fact, few countries have reached their roll-out objective of 80% smart meter coverage or higher, but partial penetration is still the norm in most distribution grids. Depending on the country's CBA, this might be a transitional phase or a permanent status. The reluctance of a minority of the population to the roll-out of smart meters and slower installation rate because of the COVID-19 pandemic must also be taken into account. In addition, measurement data and, if available, digital grid models are inevitably prone to inaccuracies, anomalies, and missing values which impact the efficiency of data-based approaches. Many case studies presented in the literature unfortunately rely on perfect synthetic measurement data and on simplistic test grids that do not reflect the situation in reality. Furthermore, besides being prone to failure, communications networks and data management systems also have limitations, notably in terms of throughput and processing capacity, respectively. This restrains the amount of data gathered within the AMI. Among other things, this impacts the number of measurement devices, the recorded quantities, and the data resolution, despite the probably greater technical capacity of the devices. Apart from smart meters and measurement devices at distribution transformers, the number of more advanced meters such as sub-metering devices and micro-PMUs is still marginal due to their relatively high cost with respect to the potential benefits. Finally, the efficiency of applications based on measurement data does not only rely on their quality and quantity but also on the acceptance and the behavior of the data owners, i.e., end consumers or prosumers. On the one hand, they have justified data protection and privacy concerns. On the other hand, they rarely take the most rational decisions when they are active stakeholders.

These various limitations obviously do not mean that measurement data, especially from smart meters, are currently not exploitable, but that they must be explicitly considered in the design of data-based approaches. On the one side, distribution system operators, aggregators, and energy providers usually lack the expertise to properly leverage a large amount of measurement data and are sometimes even not aware of the various applications. On the other side, the scientific power community should carry out data-based analysis on the basis of realistic case studies and measurement setup. Hence, this thesis intends to bridge some of the gaps which currently prevent power utilities and their customers from making use of the full potential of actual measurement data in distribution grids.

3

DATA SETS AND PREPARATION

> *If 80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team.*
>
> — Andrew Ng

Real-world AMI data sets serve as a starting point for the different applications and use cases presented in this thesis. They are described in the first part of this chapter and illustrate the diversity of data available in distribution grids. In the second part, the focus is given to the general preparation process of AMI data. Original data are inevitably prone to errors and inconsistencies, which requires some processing to increase data quality before their use for further analysis. The proposed data preparation practices are mainly based on the experience gained when processing the various real-world data sets detailed in the first instance.

3.1 INTRODUCTION

Before discussing possible applications of AMI data, it is essential to understand more concretely what are AMI data. This is the purpose of this chapter which is actually split into two main parts. In a first instance, real-world data sets are described in the frame of three different systems in Switzerland and in Costa Rica. They illustrate the status and diversity of measurement data as well as other types of data available at the level of distribution grids and of end electricity customers. It must be noted that access to such data sets is not self-evident. In this work, the data sets are part of multiple projects in collaboration with DSOs and entities responsible for data collection, and most of them are subject to a Non-Disclosure Agreement (NDA). The description of the real-world data sets and their processing is limited to non-sensitive data that are necessary to understand the context in which this work has been carried out.

In the second part, the preparation of AMI data is discussed. In fact, data recorded and communicated by each measurement device is structured according to its own standards and protocols. In a system with different types of measurement devices, measurement data must first be formatted into a common data structure. From this point, standardized cleaning methods can deal with the formatted raw data which are always prone to errors, inconsistencies, anomalies, and data gaps. Among others, algorithms for data wrangling, data quality diagnosis, outlier and anomaly detection, and missing values imputation have been developed. Part of the data cleaning methods relies on preliminary statistical analysis in order to distinguish between normal and abnormal data instances. The preparation of measurement data is closely linked with the preparation of so-called metadata which enable the transformation of measurement data into meaningful information, from which knowledge is extracted. The main data preparation steps are illustrated by examples from the real-world data sets.

3.2 PRESENTATION OF AMI DATA SETS

One of the main contributions of this work is the use of real-world data for assessing their potential for different applications. Indeed, when stakeholders in a certain power distribution system want to leverage data, they are constrained to use data available in the corresponding AMI. However, real-world data in distribution grids are typically confidential and rarely shared with power system researchers. Hence, literature on data-based studies and approaches highly relies on synthetic or simulated data which are generally not representative of actual AMI data.

Regarding distribution grid models, many publications are based on synthetic benchmark networks developed by the CIGRE Task Force C6.04.02 [212]. These grid models are integrated into the widely used *pandapower* library for power grid calculation and optimization [213]. One of the main purposes of CIGRE benchmark grids is the validation of methods for DER integration. Figure 3.1 represents the corresponding LV benchmark network. It consists of a subnetwork with 18 nodes and six residential consumers, a subnetwork with one industrial consumer, and a subnetwork with 20 nodes and eight commercial loads. More details regarding the CIGRE MV benchmark network can be found in [214]. Alternatively, IEEE distribution test feeders are often used as benchmarks for assessing (optimal) power flow methods, designing equipment placement techniques, testing islanded operations, or developing state and parameter estimation techniques [215]. These different benchmark networks are appropriate for comparison purposes among different techniques proposed in the literature. Nevertheless, most benchmark networks focus

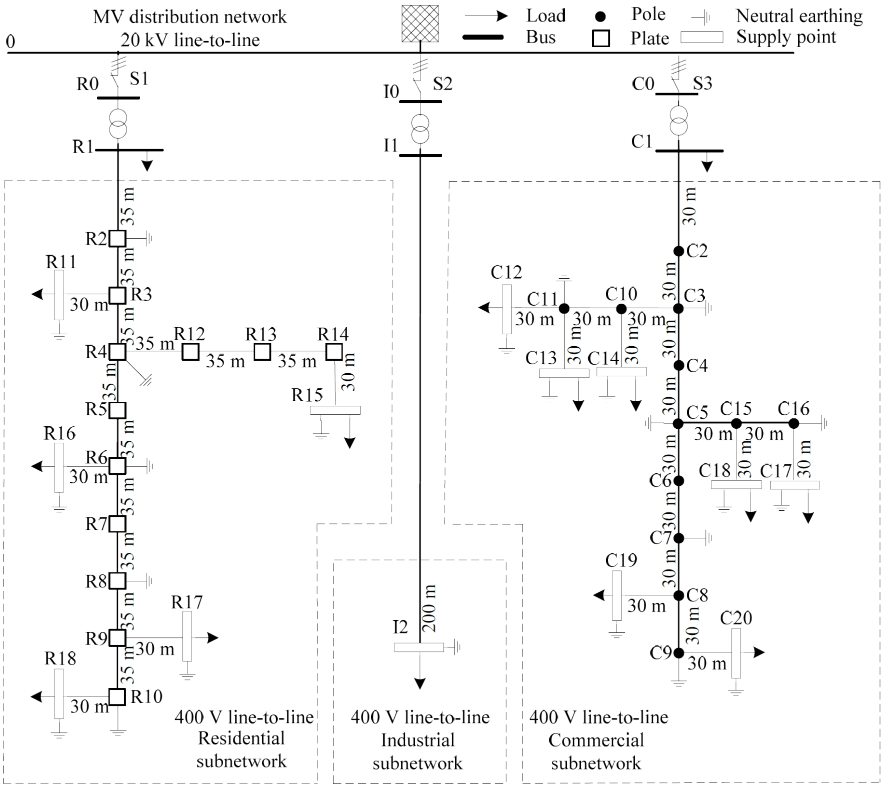


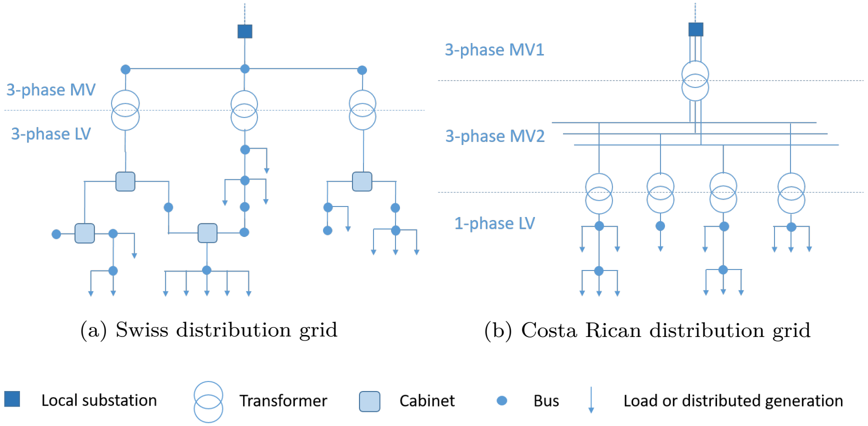
FIGURE 3.1: CIGRE benchmark model of low-voltage distribution network [213].

on the MV level, and their system design does not reflect the complexity of real distribution grids that generally consist of several hundreds or even thousands of nodes and prosumers. Limitations of currently proposed test networks are discussed by the authors in [216]. They point out the unrealistically small size of test networks, the lack of time-series data, the lack of representativeness with respect to the particular zonal characteristics of actual networks, the absence of geographical information, and the fact that test feeders are designed only for very specific problems. There is a clear need for more realistic test networks or even real-world grid models.

Regarding time-series measurements, most data refer to end-users, which raises understandable privacy concerns. In practice, they are very often not available to third parties. Traditionally, so-called Standard Load Profiles

(SLPs) are used to cope with the lack of actual load data. SLPs are load profiles specific to a certain class of consumers and represent their average behavior. Unfortunately, their shape is particularly smooth and is not representative of the high stochasticity which is observed in real load measurements at the LV level. Alternatively, different modeling tools have been developed to create synthetic load profiles. For example, LoadProfileGenerator (LPG) is a well-recognized tool for modeling residential energy consumption (i.e., electricity, gas, hot water, and cold water) [217, 218]. More precisely, load curves are generated based on a full behavior simulation of the people living in a household. The tool allows the creation of customizable residential consumers and already includes 60 predefined German households. Such tools can create realistic load profiles but generally only focus on residential loads. In addition, the process for creating a good diversity of load profiles is often time-consuming, which limits the application to large populations.

Next, the authors in [4] provide a non-exhaustive list of open load data sets. They come from real systems and have been prepared to some extent by the corresponding power company and notably pseudonymized for privacy-preserving reasons. The Electricity Smart Metering Customer Behaviour Trials conducted by the Commission for Energy Regulation in Ireland are probably the main source of smart meter data mentioned in literature [219]. It consists of active power measurements at 30-minute resolution from more than 5'000 Irish households and businesses. The trials were part of a CBA of smart meters carried out between 2009 and 2010 for the purpose of a wider national roll-out. Furthermore, Pecan Street data are a collection of historical sub-metering data gathered in approximately 1'000 households from various US cities [220]. Active power measurements of most relevant home appliances are currently available over six years with a resolution between one second and one minute. Pecan Street data are particularly suitable for NILM applications due to their high spatial and temporal resolutions. Nevertheless, open measurement data sets are still an exception and largely focus on smart meter data. Furthermore, they often only reflect a small portion of the actual data in an AMI environment. They are rarely available together with other data sets in the same system, such as the corresponding digital grid model, other quantities than power measurements, or data at higher aggregation levels. Finally, different statistical or ML-based approaches which take existing smart meter data as input can be used for the synthesis of additional data. Such approaches are described in more detail in Chapter 6.



(a) Swiss distribution grid

(b) Costa Rican distribution grid

FIGURE 3.2: Schematic representation of distribution grid structures.

Some effort is made to create more realistic case studies, but there is still a large gap to achieve the level of complexity observed in real distribution systems. The lack of representative test grids and of open real-world data sets is a clear barrier to the research in this field, and especially to the implementation of data-based approaches in real systems. Unfortunately, many simplifications and assumptions are generally made regarding the input data. This influences the credibility and applicability of data-based studies and approaches currently proposed in the literature.

This thesis intends to point out some shortcomings in current literature due to the absence of real-world data and case studies. Furthermore, it demonstrates some practical possibilities to leverage AMI data for specific applications. This is why only real-world data are used for the purpose of this work. The following sections present the three main systems and their corresponding data that have been leveraged. More precisely, Sections 3.2.1 and 3.2.2 introduce a distribution system in Switzerland and in Costa Rica, respectively. The schematic structure of their network is displayed in Figures 3.2a and 3.2b, respectively. The Swiss distribution grid is similar to the European distribution grid. Its LV part is a complex three-phase network and can be weekly meshed, although it is usually operated radially. It also consists of cable distribution cabinets with protection devices and circuit breakers that allow for topology changes. In contrast, the Costa Rican distribution grid structure is closer to the US network topology, where each MV/LV transformer is single-phase and feeds a relatively lower number of customers

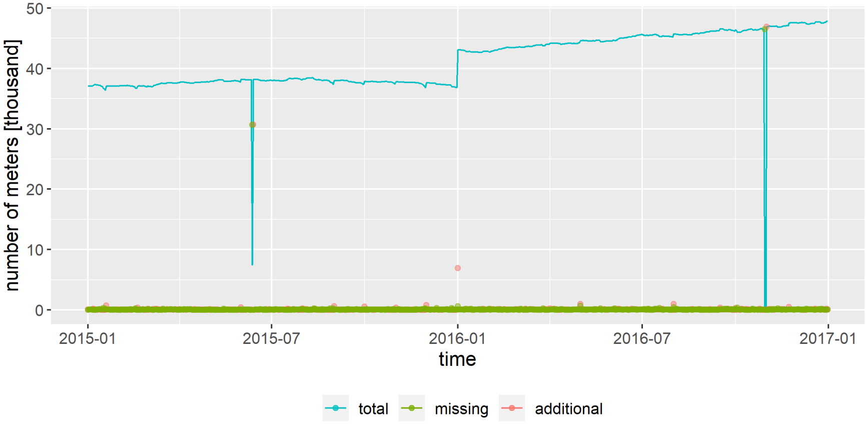


Figure 3.3: Daily roll-out statistics of smart residential meters in the City of Basel over 2015 and 2016, and number of additional and missing meter data with respect to the previous day.

(e.g., up to about 100 customers). Finally, Section 3.2.3 describes a data set of higher resolution at the end-user level.

3.2.1 *Distribution System of the City of Basel*

Industrielle Werke Basel (IWB) is the first prominent DSO in German-speaking countries to initiate a large-scale installation of smart meters [55]. IWB established a roll-out strategy well before entry into force of the Federal Energy Act that requires a smart meter penetration level of 80% by 2027. After equipping most industrial customers with AMR devices, IWB launched in 2013 the installation of smart meters for residential and commercial customers and reached a 50% penetration in 2017. Figure 3.3 illustrates the roll-out of smart meters in the City of Basel in 2015 and 2016, which covers most of the time period considered in this work. The number of active meters remains approximately constant over 2015, and more than 6'000 meters are activated on new year's day. Their number slightly increases over the second year to finally reach close to 50'000 active meters at the end of 2016. Nevertheless, most of the measurement data are not available for one day in mid-2015, whereas all data failed to be communicated over two days of October 2016. The creation and interpretation of such visualization for data preparation purposes is discussed more thoroughly in Section 3.3.2.

Access to the large database of IWB has been provided within the frame of the pilot project “Optimized Distribution Grid Operation by Utilization of Smart Metering Data” led by Adaptricity and funded by the Swiss Commission for Technology and Innovation [221]. The project’s main purpose was to investigate the potential of smart meter data in addition to traditional measurements for distribution grid planning and operation. Generally, smart meters bring more visibility into residential and commercial areas which have traditionally been considered as a black box in terms of information. Notably, smart meter measurements can be leveraged in load flow simulations to inform about the time-varying status of voltages and power flows at the LV level, assuming that a digital grid model is available. It must be noted that power flow simulations in distribution grids are commonly restricted to the MV level, where the net load of end customers is simply aggregated at the MV/LV transformers. This project led to various reflections on the realism of data-based modeling of LV grids, which is deeply discussed in Part II of this thesis.

In terms of AMI measurements, single-phase active power for about 50’000 households and commercial customers has been recorded by smart meters. Single-phase active and reactive power measurements are also available for close to 1’000 industrial customers as well as 600 PV systems. Smart meter data is sent once a day to the data management system of IWB, whereas measurements of AMR devices are gathered every month. The output temporal resolution of both smart meters and AMR devices is 15 minutes. In addition, detailed three-phase measurements (e.g., voltage, frequency, current, active and reactive power flow) at 10-minute resolution are available in several distribution cabinets equipped with the GridEye technology of depsys. Finally, the main three-phase electric quantities are also measured in local substations and in a large number of distribution transformers. In terms of communication, the distribution grid of the City of Basel was equipped with 387 Data Concentrator Units (DCUs) at the time of data preparation. In addition, each DCU was responsible for 1 to 381 valid metering devices, with an average of 70 devices per DCU.

Furthermore, a considerable portion of the distribution network of IWB has been digitized as a single-phase model in NEPLAN360, a cloud solution for grid visualization and simulation developed by NEPLAN [222]. Based on the corresponding GIS data, Figure 3.4 visualizes the topology of a large sub-grid together with connection points for loads and distributed generators in a

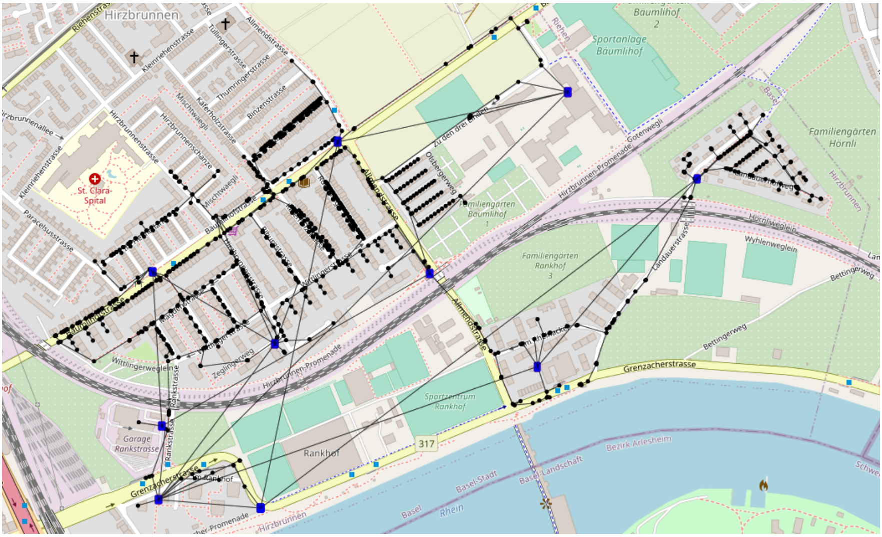


Figure 3.4: Illustration of a digitized section from the distribution network of the City of Basel operated by IWB.

mainly residential area of the City of Basel. The system is fed by one local substation and consists of 14 MV/LV distribution transformers (or feeders), 28 cable distribution cabinets, 971 buses (or connection points), from which 492 buses are loaded, and 976 lines, from which 12 lines are at the MV level (i.e., 11.7 kV). In Figure 3.4, distribution transformers and connection points are represented by blue squares and black points, respectively. MV and LV underground cables are represented by straight lines between both respective buses they are connected to. Transformers are interconnected by MV lines. For the sake of redundancy, some LV grid sections are fed by a couple of transformers normally operated in parallel. In addition, the distribution grid has a weakly meshed configuration at the LV level, and most distribution cabinets consist of circuit breakers that allow for different topologies. A connection point is always fed by only one (or a couple of) transformer(s), which can nevertheless vary depending on the configuration of circuit breakers. At the time of data preparation, the sub-grid connected 2610 residential, commercial, or small consumers, 11 industrial customers, and 17 PV systems. All commercial customers and PV systems are equipped with an advanced metering device. However, only 962 households actually had a reliable smart meter, which covered about 40% of the total residential load. This sub-grid

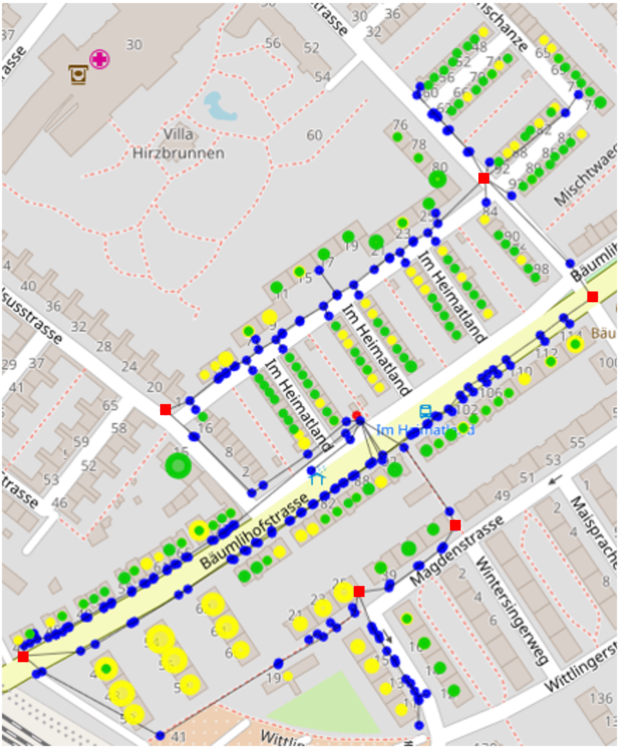


FIGURE 3.5: Illustration of a LV section with corresponding loads from the sub-grid presented in Figure 3.4.

has been chosen due to its good original data quality and availability for a study on LV grid modeling presented in Chapter 7.

Figure 3.5 zooms in on one of the LV grid sections presented in Figure 3.4. It is fed by one distribution transformer and consists of 198 power lines and 196 buses (or connection points), of which 88 buses are connected to loads and distributed generators. The feeder bus and connection points are represented by red and blue points, respectively. Moreover, red squares represent cable distribution cabinets which are located at the buses where power lines are split and at the junction with other LV grid sections. At the time of data preparation, this grid section was supplying 583 residential or small consumers, principally Multi-Family Houses (MFHs), from which 321 consumers were metered by a reliable smart meter (i.e., 55% penetration). In Figure 3.5, metered and non-metered consumers are aggregated by address and represented by yellow and green circles, respectively. The diameter of each circle is proportional to the average power consumption. Metered consumers corresponded to 58% of the total average power consumption. Moreover, six metered PV systems were installed at the top of MFHs. This grid section has been leveraged in two main studies related to pseudo-measurements synthesis and data-based preventive voltage control, presented in Chapter 6 and Section 9.5, respectively.

3.2.2 *Distribution System in Costa Rica*

Compañía Nacional de Fuerza y Luz (CNFL) is one of the main DSO of Costa Rica, providing electricity to about 500'000 customers [223]. The power company initiated its distribution grid digitalization in 2015 and has already installed about 200'000 smart meters and 3'000 data concentrator units by 2020. The main reasons for setting up an AMI are the control of non-technical losses, automated meter reading, and facilitation of billing processes. In addition, CNFL aims at automatically connecting new customers or disconnecting customers in case of non-payment as well as accessing energy consumption data on a daily basis.

Figure 3.6 illustrates the distribution grid of a neighborhood operated by CNFL in the City of San José, together with the corresponding electricity customers. MV overhead cables are represented by blue lines, and each consumer is displayed as a brown point. This sub-grid is characterized by a practically full smart meter penetration. At the time of data preparation, it connected 4634 residential loads (97.5% SM penetration), 213 commercial loads (83.6% SM penetration), and two smart metered industrial loads. In addition, the sub-grid consists of 134 MV/LV transformers feeding between 1 and 109 consumers each. Smart meter data consists of both active and reactive power measurements at 15-minute resolution. Figure 3.7 displays the active power load pattern at the substation feeding the sub-grid over a typical week[1], split by consumer type. The non-metered part represents different loads that are not measured by smart meters, as well as line and transformer losses. The neighborhood is mainly residential such that the load pattern exhibits three characteristic spikes, which coincides with the mealtime during weekdays and two consumption spikes at the weekend. PV systems are almost nonexistent.

Measurement data of this sub-grid are used for studying the effect of spatial aggregation in Chapter 5 and testing on a large scale the disaggregation

---

[1] The load pattern and magnitude are relatively similar over the whole year.

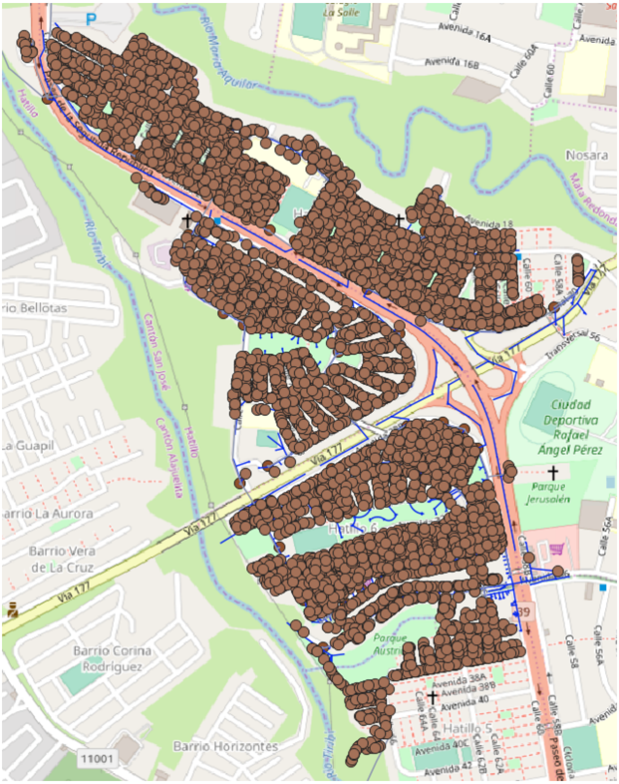


FIGURE 3.6: Illustration of a digitized section with corresponding loads from the distribution network of the City of San José operated by CNFL.

approaches in Chapter 8. The network itself is digitized as a three-phase model in openDSS, a comprehensive electric power distribution system simulator design by the Electric Power Research Institute (EPRI) [224], but has not been leveraged in this work. Data have been made available within the scope of a research visit at the University of Costa Rica in the department of power systems and electrical machines. In this thesis, this data will often be referred to as Costa Rican smart meter data set.

3.2.3 *Sub-Metering Study in Costa Rica*

Between October 2018 and February 2019, a sub-metering study has been carried out by the University of Costa Rica for Grupo ICE, a Costa Rican

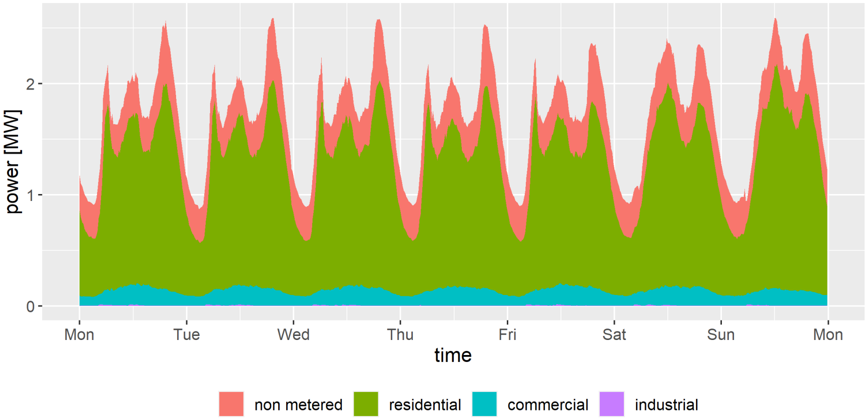


Figure 3.7: Typical weekly profile of the total consumption in the sub-grid presented in 3.6, split into residential, commercial, and industrial loads.

government-run service provider which also includes CNFL [225]. The main objective of this study is to develop and test a methodology for the determination of electricity demand curves by end-use in the residential sector. For that purpose, advanced sub-metering devices have been placed for about one week in more than 70 Costa Rican households, as represented by Figure 3.8. They measured the active power consumption of the main electrical appliances with a one-minute resolution. The measurement setup allowed a maximum of 14 channels per household.

In addition to the original study, the sub-metering data have also been used within the scope of the research visit at the University of Costa Rica. The high granularity both in terms of temporal resolution and of aggregation level (i.e., down to individual devices) is not commonplace and is of particular interest. Specifically, these measurement data have been leveraged for assessing the impact of time and spatial resolution and for the development and validation of disaggregation approaches. These two pieces of work are presented in Chapters 5 and 8, respectively.

For the purpose of this thesis, the following end-use categories have been considered: main, water heater, lighting, refrigerator, washing machine, dryer, jacuzzi, kitchen (e.g., oven, stove), others (e.g., rice cooker, microwave, coffee maker, kettle, blender, toaster, television, computer, printer, router). Water heater, refrigerator, washing machine, and dryer are treated as single cat-

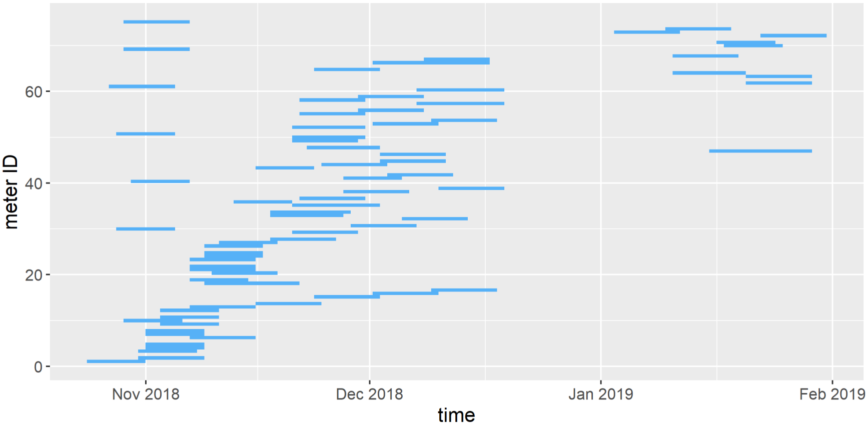


FIGURE 3.8: Representation of periods of measurement per advanced meter in the Costa Rican sub-metering study.

egories for their flexibility potential within the framework of possible DR programs. The “main” category corresponds to the total household load. The difference between the main load and the aggregation of all metered appliances is categorized as “not measured”. In addition, it comes out from the metadata that the refrigerator and the washing machine are sometimes metered on the same channel as other appliances. In some cases, metadata also indicate the presence of a non-metered dryer. This is taken into account when setting up the categories in order to keep track of their load. After data preparation, measurement data of 70 households are considered of good quality.

Figure 3.9 illustrates the load profiling over one day of one of the households under study. It appears that the water heater is responsible for substantial power consumption spikes in the morning. The dryer accounts for the majority of the activity in the afternoon and in the early evening. Furthermore, the activity of non-measured appliances is visible in the main load, especially in the early morning and in the evening. Other non-specified appliances build the base load. Over the entire measurement period, Figure 3.11a indicates that 60% of the energy consumption of this household comes from the water heater and the dryer, and about 30% cannot be specifically identified. At the level of individual households, the load is characterized by particularly high

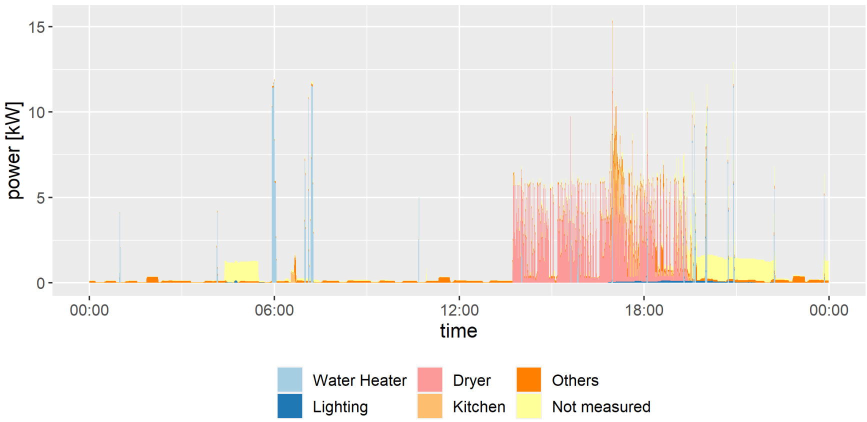


FIGURE 3.9: Load profiling of a single residential user over one day.

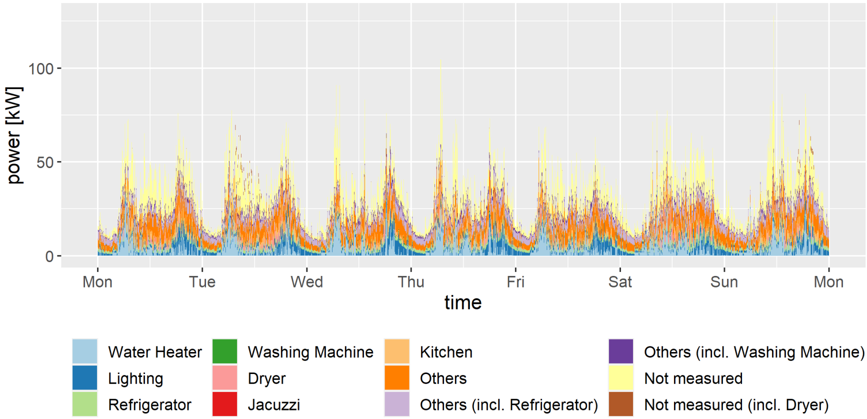
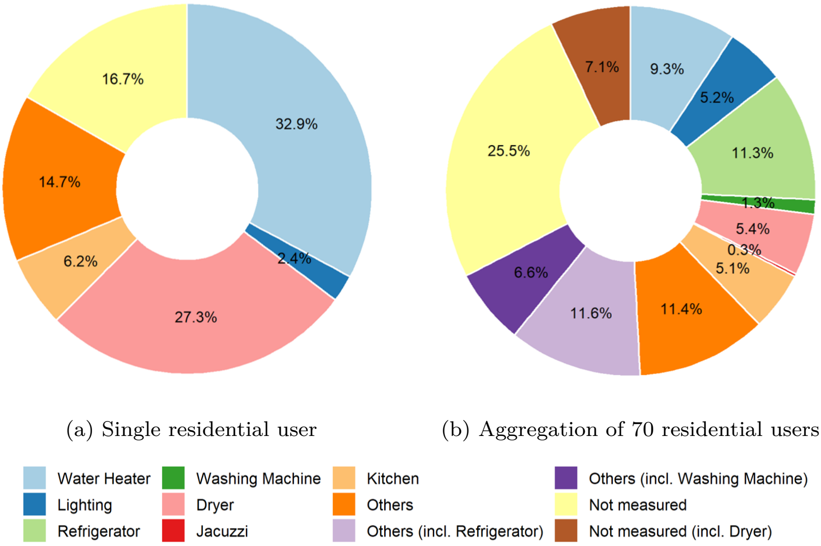


FIGURE 3.10: Load profiling of the aggregation of 70 residential users over one week.

volatility, and its pattern greatly varies between days and between different households depending on the devices in operation.

Although each household has not been recorded over the exact same time period, all load profiles of good quality have been aggregated by weekday to create Figure 3.10. This provides some intuition about load profiling at an aggregate level. The load is still volatile, but certain trends are emerging.



(a) Single residential user

(b) Aggregation of 70 residential users

FIGURE 3.11: Consumption share among residential appliances in the Costa Rican sub-metering study.

First, the main load profile follows the typical weekly pattern displayed in Figure 3.6. In addition, water heaters tend to consume mainly in the first half of the day with further smaller activity in the evening. Lighting is obviously mainly active in the evening. The aggregate consumption of refrigerators appears to be relatively constant, but a non-negligible part is metered together with other appliances. In fact, a large share of the load cannot be assigned to specific appliances. According to Figure 3.11b, this share amounts to 67%, although a substantial part corresponds to the refrigerator and the washing machine. Water heaters and refrigerators are the identified appliances with the highest share of consumption.

3.3 DATA PREPARATION

Data in distribution grids originate from a large variety of sources and rely on very diverse specifications. On the one hand, time-series measurements largely contribute to the total volume of data. They are generated by measurement devices and comprise various quantities at different locations. In distribution grids, measurement data typically come from electricity consumers (e.g., residential, commercial, and industrial customers) but are also recorded at higher aggregation levels (e.g., cable distribution cabinets, transformers, and substations) or at the level of individual appliances (e.g., PV systems, home batteries, EV chargers, HVAC and DHW systems). Depending on the needs, measurement data are recorded at different temporal resolutions and accuracy standards. In addition to purely electricity-related quantities (e.g., active and reactive power, voltage, current), exogenous data are often of interest when they influence electricity quantities. Specifically, meteorological measurement and forecast data help for explaining and predicting the behavior of electricity quantities. On the other hand, so-called metadata, which basically give information about other data sets, must also be considered. Among others, metadata inform about the characteristics of electricity consumers and producers (e.g., consumer type, building features, billing data, installed PV and battery capacity) and provide identification numbers (IDs) to link multiple data sets (e.g., link meters and customers, meters and DCUs, customers and grid buses). Finally, AMI data might also consist of digital grid models which directly describe distribution networks (e.g., line and transformer parameters, topology, connection points).

Real-world data sets are rarely of perfect quality. For different reasons, raw data sets are subject to various types of errors and inconsistencies which degrade the data quality. In the context of AMI data, common examples of data quality issues are inconsistencies between measurement data sets, inconsistencies with metadata, duplicate time-series measurements, gaps in measurements, unrealistic measurement profiles, incorrect measurement signs, incomplete or wrong list of electricity customers, etc. There is the need for a consistent procedure that deals with the most common data quality issues. Moreover, even if the original data is of excellent quality, the multiple data sets in a system come from different sources with their specific structure and standards. Further analysis and integration of the different data sets require some formatting and standardization. The process of cleaning and transforming raw data before its use for further applications is called data preparation.

Literature on the preparation of AMI data is particularly scarce. This is explained by the fact that data are essentially gathered by power companies which simply do not reveal how the data are processed for obvious confidentiality reasons. In the case of open data sources, a certain level of data wrangling and cleaning is usually already performed upstream to facilitate the subsequent processing for data or power system engineers. Anomalies and missing data are among the issues which might still not been addressed in open data sets, which explains the higher number of publications on these aspects.

Measurement data sets are generally too large to allow manual processing and visualization of individual samples. Appropriate data manipulation tools and capabilities are indispensable. For the purpose of this work, all pieces of code, functions, and algorithms for data preparation have been implemented in R, a free software environment for statistical computing and graphics [226]. This high-level language is particularly easy to learn and write, and has been specifically designed for data manipulation. RStudio Desktop has been chosen as Integrated Development Environment (IDE) for programming in R [227]. The open-source software includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging, and workspace management. Although various data cleaning and missing value imputation packages for R exist, these are usually too generic and not adapted to the often more advanced needs in this work. Hence, multiple packages, functions, and algorithms have been specifically designed and coded to address the requirements for processing time-series data from distribution grids. The different pieces of code still take advantage of some of the functions in already existing packages to speed up the manipulation of large data sets and allow for proper data visualization. The most important dependencies of the packages created for preprocessing purposes are the followings:

- *data.table* [228]: enhanced version of *data.frames* (i.e., standard data structure for storing data in R) that enables faster manipulation of large data sets
- *doSNOW* [229]: support for parallel computation
- *dplyr* [230]: grammar of data manipulation functions such as *mutate*, *select*, *filter*, *summarise*, *arrange*, and *group_by*
- *ggplot2* [231]: popular library for the creation of graphics based on *The Grammar of Graphics* [232]
- *lubridate* [233]: set of functions to manipulate date-time objects
- *plotly* [234]: library for the creation of interactive, publication-quality graphics
- *reshape2* [235]: set of functions that facilitate data transformation between wide and long formats

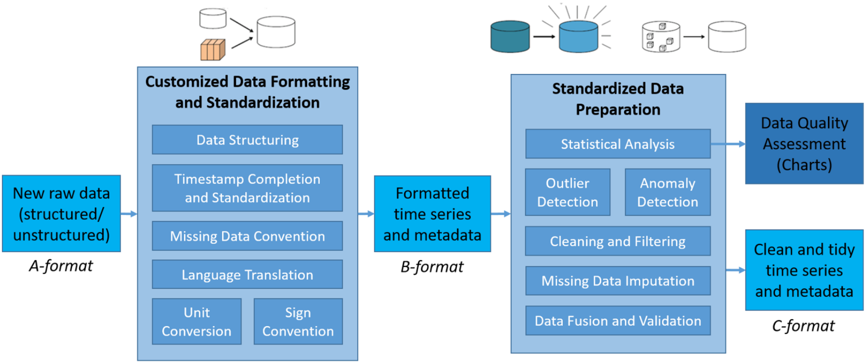

New raw data (structured/ unstructured)
A-format
Customized Data Formatting and Standardization
Data Structuring
Timestamp Completion and Standardization
Missing Data Convention
Language Translation
Unit Conversion
Sign Convention
Formatted time series and metadata
B-format
Standardized Data Preparation
Statistical Analysis
Outlier Detection
Anomaly Detection
Cleaning and Filtering
Missing Data Imputation
Data Fusion and Validation
Data Quality Assessment (Charts)
Clean and tidy time series and metadata
C-format

FIGURE 3.12: General pipeline for data preparation, transforming raw data into clean and tidy time series and metadata.

Figure 3.12 illustrates the general preparation pipeline followed in this work in order to prepare the various raw data sets for further analysis. The starting point of this pipeline is raw data which basically consist of structured and unstructured data. Structured data represents the largest volume of data available in the AMI environment and can be further split into time-series measurement data, which are produced by measurement devices, and structured metadata, which mostly refer to lists of electricity customers, buildings, devices, or connection points with their individual characteristics. The digital model of a grid is also considered structured data. Structured data can be stored in a database but also in files, e.g., in Comma-Separated Values (CSV) or Extensible Markup Language (XML) files. It must be noted that measurement data are largely generated automatically, whereas metadata are partially created manually. Errors tend to be systematic in measurement data in contrast to metadata. In addition, there are so-called unstructured data that have no predefined format or organization. Such data usually consist of different pieces of information that help data engineers to make sense of the structured data, like *readme*-files, pictures, reports, side notes, etc.

Due to the large variety of data sources and formats, these raw data cannot be directly processed following a standardized procedure. They require customized wrangling algorithms to comply with predefined data structures and standards, as detailed in Section 3.3.1. In the next step, formatted time series and metadata are subject to statistical analysis for data quality diagnosis, which is detailed in Section 3.3.2. Such statistical analysis serves as the basis for data cleaning and filtering following a standardized procedure. This is explained in detail in Section 3.3.4 that addresses the detection and correction of anomalies and inconsistencies, and the removal of data of particularly poor quality. In addition, Section 3.3.3 defines how point outliers are detected and treated. Furthermore, all raw data sets certainly contain missing values, which can be handled by imputation algorithms, as detailed in Section 3.3.5. Finally, Section 3.3.6 details how multiple clean and tidy data sets are potentially compared and even merged with the aim of further improving the data quality, consistency, and comprehensiveness. In this work, A-format refers to the original raw data, B-format refers to standardized and formatted data before cleaning, and C-format refers to clean and tidy data at the end of the preparation process.

The following sections principally focus on the preparation of time series measurement data. The preparation process is illustrated by examples taken from the different data sets used in this work. For the sake of conciseness, only the main issues that have been observed in AMI data sets are discussed. In fact, the preparation and quality assessment of real-world data sets are more exhaustive than the examples that are presented. Each data set has its own specifications and data quality issues that require substantial effort and resources to deal with. In data analytics, it is commonly recognized that 80% of the work is dedicated to data preparation, whereas only the remaining 20% focuses on actual data analysis. Special care must also be given to the preprocessing of metadata, but the main steps are analogous to the preparation of measurement data and are not explicitly described. In addition, a similar process is necessary for the preparation of grid models. Their digital version is essentially based on hand-written notes and diagrams, and is inevitably prone to various errors. Nevertheless, the preparation process of grid models is out of the scope of this work.

3.3.1 *Data Standardization and Formatting*

Standards and specifications regarding the representation of each piece of data can significantly vary among data sets coming from different systems. First of all, timestamps, which are associated with data points of a time series, can be encoded in various ways and might refer to different time zones. For the purpose of this work, timestamps are converted to Coordinated Universal Time (UTC) and encoded according to ISO 8061, e.g., 2021-01-26T07:52:57Z. Specific attention must be given to changes between Daylight Saving Time (DST) and standard time. Next, missing data can typically be represented by

Quantity	Nomenclature	Number of decimal digits	Unit
Active power	p_1, p_2, p_3, p_tot, p_avg	0	W
Reactive power	q_1, q_2, q_3, q_tot, q_avg	0	Var
Voltage	u_1, u_2, u_3, u_avg	1	V
Current	i_1, i_2, i_3, i_tot, i_avg	2	A
Power factor	cosphi_1, cosphi_2, cosphi_3, cosphi_avg	2	-

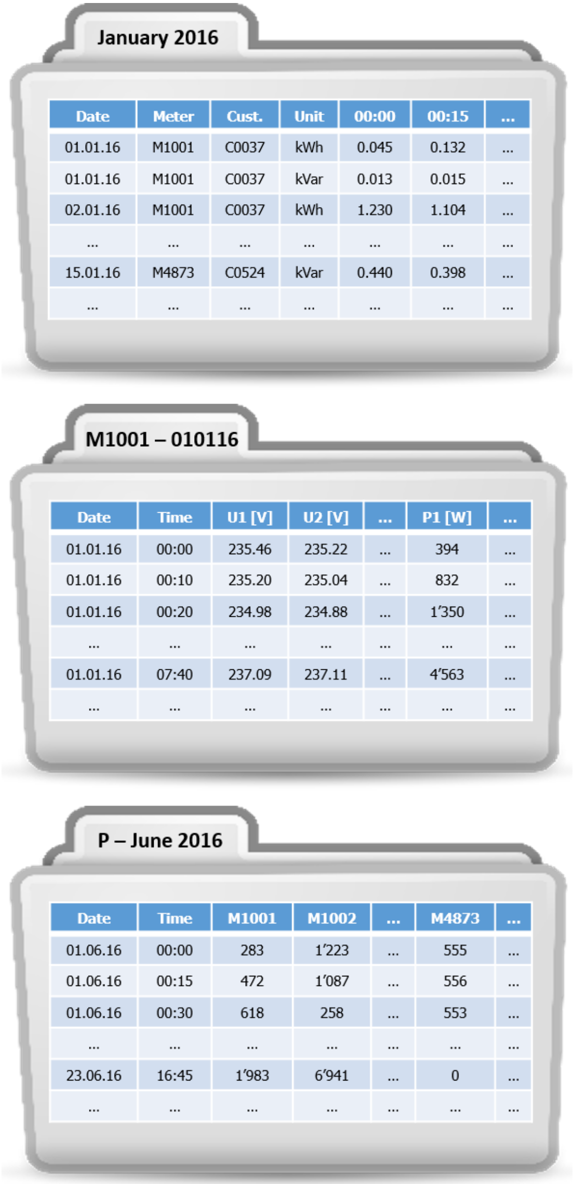
TABLE 3.1: Specifications of most common measurement data types in B-format.

the symbols NA (not available) or NAN (not a number), but sometimes also by a value of zero, by the repetition of the latest measured value, or by the absence of value. Luckily, the representation of missing data within a data set is usually consistent. In this work, missing data are encoded as NA, which is identifiable by all data processing tools. It is also common to notice the absence of certain timestamps with respect to the assumed temporal resolution. In any case, all recorded timestamps are ordered, and missing timestamps are completed while the corresponding measurement values are defined as NA. Next, the language of data, and especially metadata, essentially depends on the region where they come from. For the sake of consistency, all languages are translated to English in this work. Furthermore, the nomenclature of the different quantities, the number of decimal digits, the unit of measurement as well as the sign convention must be standardized. Table 3.1 summarizes the specifications for the most common quantities. It must be noted that energy values are converted into power values, which facilitates further transformations, and especially modifications of the temporal resolution. In terms of nomenclature, quantities are represented by their usual symbol followed by the indication of the phase or whether it refers to the sum of the three phases or to the average value. The number of decimal digits is a trade-off between the desired accuracy with respect to measurement data in LV grids and the storage requirements. Finally, active power and current are defined as positive when they are consumed, and reactive power is defined as positive when it is inductive. Other quantities can also be represented in a similar way, like more advanced measurements (e.g., grid frequency, total harmonic distortion) or the binary status of devices (e.g., opening of circuit breakers, availability of a controllable device).

In addition, very diverse layouts can be observed for raw AMI data sets depending on their source. When stored in a database, its architecture characterizes the data structuring and allows for a certain level of versatility. Databases are nevertheless out of the scope of this section, and discussions on data representations primarily focus on their storage in files. In most cases, AMI data are represented in tabular form (i.e., with rows and columns), where CSV and TXT files are the most common formats. In the case of time-series measurements, original data sets may contain information stemming from various sensors, quantities, and time steps, among others. Knowing that these different aspects must be structured based on the two dimensions inherent to the tabular form and possibly across multiple files, this leads to a large variety of options.

Figure 3.13 illustrates three different layouts of measurement data in tabular form that have been observed among the data sets presented in Section 3.2. In the first example, each file corresponds to a month and consists of active and reactive energy measurements from various meters and customers. Each row contains the measurements over a full day for a specific meter, a specific customer, and a specific quantity. In the second example, each file corresponds to the measurements of a specific meter on a specific day. The various quantities (e.g., voltage at phase 1, voltage at phase 2, active power at phase 1) are displayed in different columns, whereas each row refers to a specific time of the day. In the third example, only one quantity (e.g., active power) for a specific month is stored per file. Measurements of each meter are displayed in a separate column, whereas the different time steps of the month correspond to the different rows. All files have in common the recording of measurement data for a limited period of time, e.g., one month or one day. This is explained by the fact that measurement data are generally sent on a periodic basis by advanced meters to the central data management system. Hence, the frequency of data pooling determines the length of the time period considered in each raw data set.

Moreover, some pieces of metadata might be included within the file of measurement data, usually in dedicated columns, such as the customer address and/or name, connection point to the grid, associated DCU, unit of



January 2016

Date	Meter	Cust.	Unit	00:00	00:15	...
01.01.16	M1001	C0037	kWh	0.045	0.132	...
01.01.16	M1001	C0037	kVar	0.013	0.015	...
02.01.16	M1001	C0037	kWh	1.230	1.104	...
...	...	...	...	...	...	...
15.01.16	M4873	C0524	kVar	0.440	0.398	...
...	...	...	...	...	...	...

M1001 – 010116

Date	Time	U1 [V]	U2 [V]	...	P1 [W]	...
01.01.16	00:00	235.46	235.22	...	394	...
01.01.16	00:10	235.20	235.04	...	832	...
01.01.16	00:20	234.98	234.88	...	1’350	...
...	...	...	...	...	...	...
01.01.16	07:40	237.09	237.11	...	4’563	...
...	...	...	...	...	...	...

P – June 2016

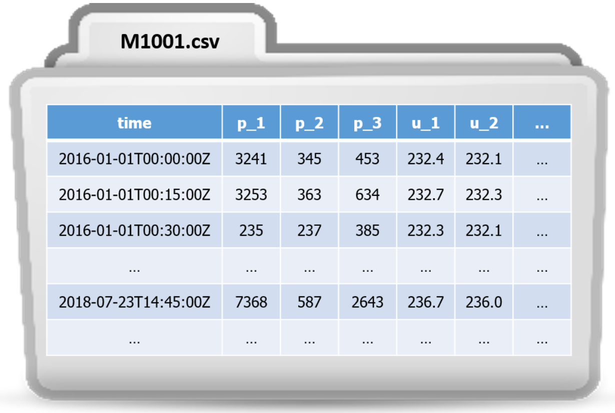
Date	Time	M1001	M1002	...	M4873	...
01.06.16	00:00	283	1’223	...	555	...
01.06.16	00:15	472	1’087	...	556	...
01.06.16	00:30	618	258	...	553	...
...	...	...	...	...	...	...
23.06.16	16:45	1’983	6’941	...	0	...
...	...	...	...	...	...	...

FIGURE 3.13: Example of different original structures of AMI measurement data sets in tabular form.

measurement, etc. Alternatively, time-series measurement data can be provided together with some metadata in a tree-like structure such as the XML format. In general, most of the metadata are still provided in separate data sets and consist of very diverse information which must also be formatted, filtered, and cleaned.

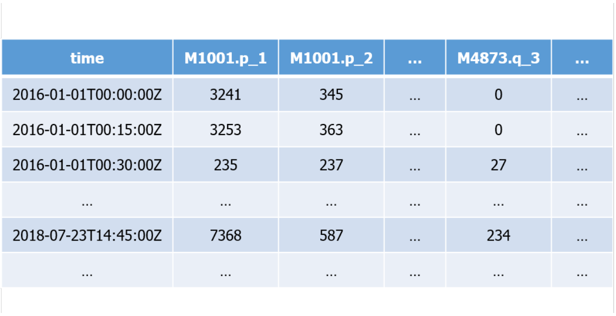
For the purpose of this work, the structure of raw measurement data and metadata is standardized into the so-called B-format which has been developed with the help of data engineers at Adaptricity. In this case, all measurement data related to a specific meter are stored as a unique data set in one file whose name corresponds to the meter ID. In this way, the size of each file remains manageable and does not depend on the number of meters but only on the number of measured quantities, temporal resolution, and total length of the measurement period. Figure 3.14a provides an example of measurement data in B-format when stored in a file. The first column corresponds to the timestamps in ISO 8061 standard, whereas the following columns consist of measurement data of different quantities according to the specifications mentioned in Table 3.1. When it comes to processing measurements from different meters, the multiple data sets are merged together into a modified B-format as illustrated in Figure 3.14b. The main difference with the original B-format is that measurements of different meters are stored across multiple columns. This implies that all measurements must comply with the same timestamps and might necessitate the addition of leading and trailing NA values to some time series. The terminology of column names consists of both a meter ID and a quantity, separated by a dot.

In R, such data sets can be converted into data tables as defined by the data.table package, which allows for more efficient data manipulation and speeds up data reshaping as well as the selection and filtering of subsets. Hence, a limited time period or a subgroup of quantities or meters can be rapidly extracted from large data sets. Specific functions have been designed in R to support data manipulation in B-format and import or store data from or to CSV files, respectively. Regarding metadata, only one file per system is usually created, which contains all meter IDs in the first column and all corresponding pieces of information in the following columns. Each meter ID must be unique and appears both in the formatted metadata file and as the name of the corresponding measurement data file. Preparation of measurement data and metadata is typically carried out simultaneously since the original data sets might consist of both types of data.



time	p_1	p_2	p_3	u_1	u_2	...
2016-01-01T00:00:00Z	3241	345	453	232.4	232.1	...
2016-01-01T00:15:00Z	3253	363	634	232.7	232.3	...
2016-01-01T00:30:00Z	235	237	385	232.3	232.1	...
...	...	...	...	...	...	...
2018-07-23T14:45:00Z	7368	587	2643	236.7	236.0	...
...	...	...	...	...	...	...

(a) Storage in file



time	M1001.p_1	M1001.p_2	...	M4873.q_3	...
2016-01-01T00:00:00Z	3241	345	...	0	...
2016-01-01T00:15:00Z	3253	363	...	0	...
2016-01-01T00:30:00Z	235	237	...	27	...
...	...	...	...	...	...
2018-07-23T14:45:00Z	7368	587	...	234	...
...	...	...	...	...	...

(b) Storage in data manipulation tool

FIGURE 3.14: Illustration of measurement data set in B-format.

Therefore, customized converters must be designed to ensure that the different pieces of information follow the same standards and to format each new raw data set into a common structure. This part of the data preparation process is particularly time-consuming since it requires studying the assumed conventions and understanding how the data have been originally structured before writing new customized pieces of code and routines for data wrangling. An additional function of these converters consists of merging new data sets with already formatted data sets. Notably, this concerns the integration of a new day or month of measurement data, or the addition of new electricity customers to the existing list of customers. New measurement data sets might require large storage capacity, especially when the number of meters is substantial, and might be difficult to process all at once due to memory limitations if the preparation is performed on a single local machine. On the one hand, converters should be able to quickly detect and discard irrelevant pieces of information like duplicate and NA time series, redundant metadata, or quantities that are not of interest. Exact duplicates might appear when a meter transmits its data via multiple DCUs. When it comes to storing measurement data of individual meters into separate files, trimming leading and trailing NA values allows for size reduction without loss of information. On the other hand, converters might need to split the original data set into multiple chunks, ideally corresponding to subgroups of meters, that are then standardized separately.

In addition, metadata or unstructured data help for the integration of new data sets. Handling changing customer or meter IDs in subsequent data sets is a common example. Unstructured data are particularly valuable to provide information or explain atypical issues such as incorrect or missing IDs, large data gaps, and abnormal measurement data. It should be noted that customized converters are also required for digital grid models, although formats and standards seem to be more unified in comparison with measurement data and metadata. They are often available in XML and JSON formats but also in formats specific to grid simulation and visualization tools like openDSS [236] or Neplan [222]. Digital grid models might also include GIS features that allow the visualization of its structure on a map.

3.3.2 *Statistical Analysis*

Standardized data preparation begins with a statistical analysis which is carried out first at the macro-level and then at the level of individual time series. In contrast to the pure data standardization and formatting, this stage aims to make sense of the raw data and initiates its transformation into valuable information.

First, roll-out statistics are performed based on the meter IDs available in the original data sets, as illustrated by Figure 3.3 for the case of the City of Basel. As previously mentioned, each original data set typically corresponds to a limited period, typically one day. For each day, the total number of active meters, as well as the number of additional and missing meters with respect to the previous day, are calculated, which can be conveniently performed and updated during the integration of new data sets. Among others, this provides insight into the number of functional meters with respect to the theoretical number of installed meters, and by extension, into the number of defective devices. A drop in the total number of meters for a certain period indicates a probable data communication or storage issue. Moreover, relatively large numbers of additional and missing meter IDs for a certain day without substantial variation in the total number might reflect a modification of meter IDs. Similar analysis can be performed regarding the overall proportion of missing data per day, where a relatively high percentage for a certain period would reflect temporary data communication or storage issues.

Due to the usually large number of samples in measurement data sets, individual inspection and processing of each time series is not an option. Hence, different pieces of information can be extracted, and statistical measures can be computed to summarize each time series into key features, which helps for data quality diagnosis. Depending on the usage, statistical measures can typically be computed either over one year or over the entire available measurement period. This builds metadata which are stored in a specific file where each row refers to a different meter and quantity. Here are the main pieces of information and statistical measures considered in this work:

- First and last timestamps with non-NA value: This informs about the bounds of the time period with actual measurements.
- Mean and median values: Basic statistical measures that provide insight into the central location of the data.
- Minimum and maximum values: Basic statistical measures that define the range of recorded values.
- Number and percentage of missing values (NA): These measures obviously inform about the data quality.
- Number and percentage of zero values: In some original data sets, missing data are encoded as zero values such that an abnormally high percentage of zero values in a time series could reveal bad data quality.
- Average and maximum length of a sequence of missing values: Beyond the sole percentage of missing data, this gives an indication about the type, and by extension, about the reasons for missing data. Especially, it indicates whether the time series contains large gaps or whether missing data are dispersed over the measurement period. It must be noted that leading and trailing NA values are not taken into account.
- Average and maximum length of a sequence of constant values (zero-order hold): In some original data sets, missing data are encoded as zero-order hold values. In other words, in the case of a gap in measurements,

recorded values might be defined as equal to the last measured value until a new actual measurement is recorded.

- Autocorrelation value with daily and weekly lags: Electrical measurements tend to have a certain periodicity, particularly with a cycle of one day and one week, which can be quantified by the autocorrelation function at the desired lag.
- Correlation with exogenous profiles: Electrical measurements, notably active power consumption and production, might be influenced by exogenous factors such as meteorological conditions. Hence, an unexpected Pearson correlation coefficient of electric quantities with supposedly influencing factors is a clue for abnormal data.
- Additional statistical measures: Depending on the usage of the data set, further statistical measures might be needed (e.g., variance, percentile, probability distribution).

Visualization of those features helps data analysts to make sense of time-series measurements. For example, the detection of the first and last timestamps has been leveraged in Figure 3.8 to present the different periods of measurements among the advanced meters in the Costa Rican sub-metering study. In addition, the distribution of calculated statistical measures can be visualized under the form of a histogram, which provides insight into the quality of the data set. For the purpose of this section, data sets from the City of Basel are chosen as an example. Figure 3.15 illustrates the percentage of missing values in the active power measurements of all smart metered residential consumers. In that respect, it comes out that the data set is generally of good quality and a large majority of time series are practically free of missing values. The number of time series with an increasing percentage of missing values exponentially decays. Analogously, Figure 3.16 exhibits the Pearson's correlation coefficient between the active power production of PV systems and the solar irradiance measured by MeteoSwiss in the City of Basel [237]. It must be reminded that the load convention applies to active power measurements such that PV production is recorded as negative data. As expected, most of the time series are highly correlated with solar irradiance. Slightly lower correlation coefficients in absolute terms might be explained by the effect of local shading or by the fact that solar panels might not be directly close to the weather station, which leads to significant discrepancies during partly cloudy days. Nevertheless, it also appears that a few samples exhibit a clearly abnormal correlation coefficient (i.e., too close

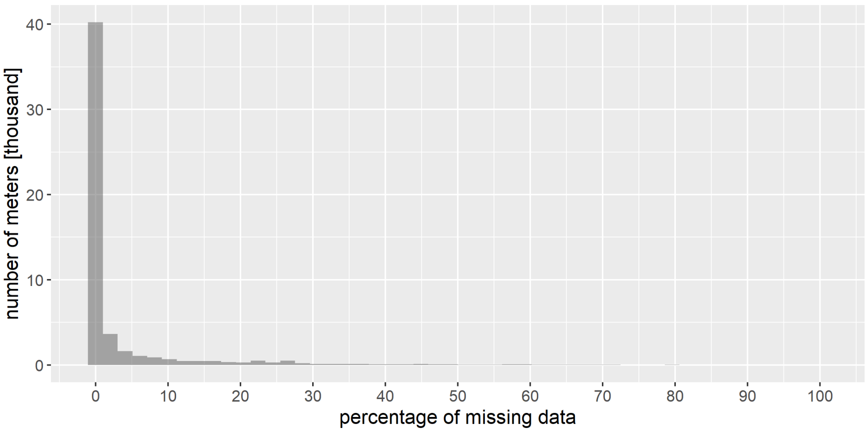


Figure 3.15: Histogram of the percentage of missing values in active power measurements of residential consumers in the City of Basel.

to zero or even positive). This needs to be further investigated, notably by visualizing the time series in question. It turns out that most abnormal time series in this data set do not correspond to the production of PV systems. A similar analysis is performed with all relevant statistical measures, which is however disregarded in this section for the sake of conciseness.

3.3.3 *Outlier Detection*

Broadly speaking, outliers are individual data points that differ significantly from the majority in a data set. They must be detected during the preparation of time-series measurements since they can be a sign of bad data. In statistical terms, a data point is defined as an outlier if it lies well above the third quartile (i.e., 75th percentile) or well below the first quartile (i.e., 25th percentile) [238]:

$$y_i = \text{outlier} \iff \begin{cases} y_i > Q3 + \alpha \cdot (Q3 - Q1) & \text{or} \\ y_i < Q1 - \alpha \cdot (Q3 - Q1), \end{cases} \tag{3.1}$$

where y_i is data point i in a certain data set, and $Q1$ and $Q3$ are the first and third quartiles, respectively. The difference between $Q3$ and $Q1$ is defined as the Interquartile Range (IQR), where the middle 50% of data points lie. α is a factor that determines how far from the IQR a data point should be in order to be defined as an outlier. In the literature, α is commonly set to 1.5

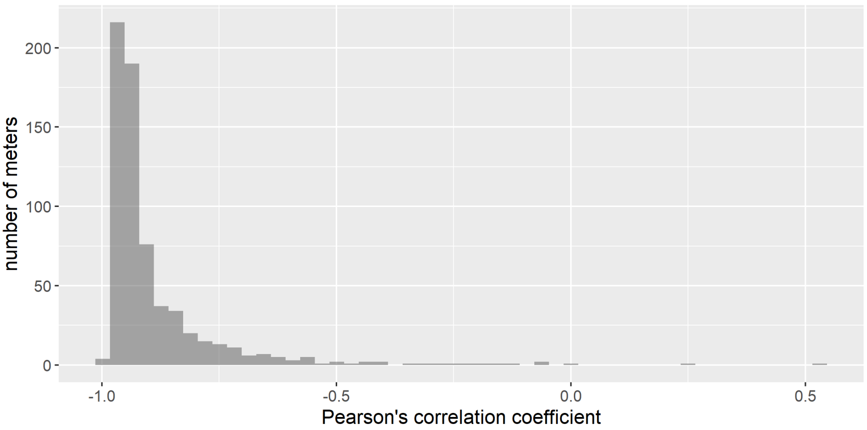


FIGURE 3.16: Histogram of the Pearson's correlation coefficient between the solar irradiance and the active power production of metered PV systems in the City of Basel.

for weak outliers and to 3 for strong outliers. When it comes to detecting bad data in AMI measurements, the value of α must be adapted depending on the type of data. For relatively smooth data such as the power loading of transformers or voltage measurements, $\alpha := 3$ is totally appropriate. However, many AMI data are highly volatile such that α must be set to much higher values to avoid unnecessary detection of outliers that turn out to be valid data. In the case of active and reactive power measurements at the consumer level, $\alpha := 10$ is a reasonable choice. For example, Figure 3.17 illustrates the detection of outliers in the active load profile of a residential consumer. All outliers are in order of magnitude (i.e., MW) much greater than what is realistically assumed for such load and indeed correspond to bad data. In such a case, outliers are replaced by NA values which are handled together with other NA values during the missing data imputation stage (see Section 3.3.5).

Nevertheless, it happens that data points are defined as outliers, although they are valid measurements. This can be the case for certain individual devices which are on standby mode most of the time and consume substantially high power when active (e.g., Domestic Hot Water (DHW) systems, tumble dryers, ovens). Due to the low usage rate, the IQR is particularly small, and the activity of such devices translates into outliers. Therefore, the detection of outliers in a time series necessitates a more thorough inspection. Depending

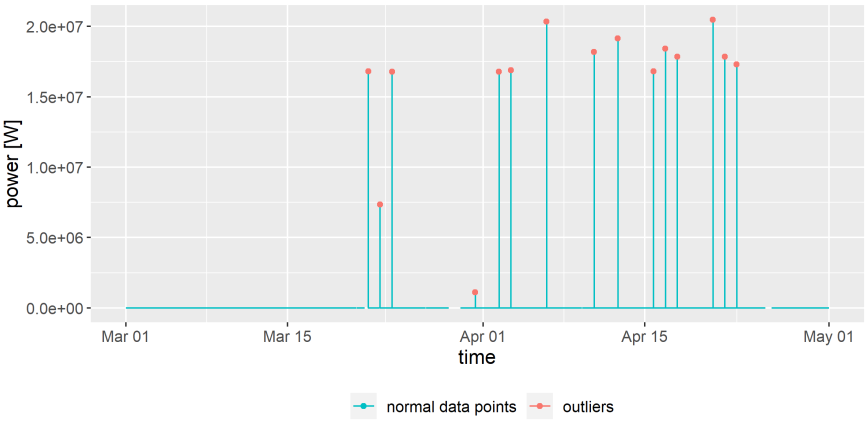


FIGURE 3.17: Example of a time series with active power measurements where outliers are detected ($\alpha := 10$).

on the type of measurement, the definition of an outlier might need to be adapted, e.g., by relying on a wider range than the IQR.

3.3.4 *Anomaly Detection, Filtering and Cleaning*

The detection and handling of anomalies in time series is an important step of data preparation. Literally, the term “anomaly” is very similar to an “outlier”, although an anomaly suggests that it differs from certain assumptions that are made about the normal behavior of the data. Hence, its definition is more subjective and particularly broad. As for outliers, anomalies are a sign of bad data but do not necessarily mean bad data, such that closer inspection is necessary. Basically, anomalies can be divided into the following subgroups [239]:

- Point anomaly: Unique data point which is outside of the usual range of data points in the time series.
- Abnormal sequence of data points (or collective anomaly): Sequence of multiple data points in a time series whose joint behavior is unusual, although each individual observation is not necessarily a point outlier.
- Abnormal time series: Entire time series whose shape or behavior considerably varies from the majority of time series in the same data set.

Point anomalies are essentially point outliers as discussed in Section 3.3.3. Here, the focus is on abnormal sequences of data points in a time series and on abnormal time series in a data set. There is extensive literature on anomaly detection for time series and increasing interest in the context of AMI data. This section does not intend to provide an exhaustive literature review but mainly to describe what are the most commonly observed anomalies in AMI data and how they are handled for the purpose of this work. It must also be specified that Anomaly Detection Techniques (ADTs) proposed in the literature are usually too generic and cannot be blindly applied to AMI measurement data. Especially, they might be highly volatile and do not necessarily follow a known distribution which is often a prerequisite for generic ADTs. In this context, domain knowledge is crucial to take into account the data particularities and to focus on specific indicators of bad data quality.

When it comes to detecting abnormal sequences of data points in a time series, supervised learning algorithms such as ARIMA and PARX models, SVRs, RNNs, LSTMs, and autoencoders are often suggested in literature [240–242]. They basically work as forecasters and are trained to estimate the normal behavior of the data. Hence, a sequence of actual data points is defined as abnormal as soon as its difference with the prediction exceeds a certain threshold. These techniques can also be adapted to online anomaly detection. Nevertheless, the detection of anomalies highly depends on the design of such models and on the selected threshold. In addition, their training process is particularly time-consuming, especially accounting for the large number and variety of time series measurements gathered in AMI. Alternatively, clustering-based and entropy-based algorithms are proposed in literature [243, 244]. While complex anomalies due to intentional data manipulation, e.g., for the purpose of energy theft, definitely require advanced anomaly detection techniques, this is out of the scope of this work. From experience, it has been noticed that anomalies are often not malicious but can still appear in very various forms. They can originate from any stage between measurement by advanced meters and storage in the data management system.

First, an abnormally high or low sequence of data points can already be identified by the outlier detection approach detailed in Section 3.3.3. Next, it happens that missing data are encoded as a sequence of constant values or zero values, as mentioned in Section 3.3.2. Based on the statistical analysis, relatively long sequences of constant values or an abnormally high percentage of zero values in a time series with respect to other time series of the same data set are a clue of missing data. These features can be visualized similarly to Figures 3.15 and 3.16. The threshold that distinguishes between abnormal and normal features highly depends on the type of data. For example, in the case of PV production, long sequences of zero values at night totally make sense, but some production during the day should be visible, even for cloudy days, such that the threshold can reasonably be set to one day. For electric devices following an ON-OFF controller (e.g., DHW system, heat pump, refrigerator), sequences of constant power values are normal, and the threshold must be considerably larger than the usual duration of ON and OFF periods. However, the load of residential or commercial consumers is supposed to be volatile and such phenomena over long time periods are abnormal. On this basis, holidays should not be detected as abnormal since some electric appliances like the refrigerator must still work and devices on standby mode still show some activity.

Moreover, different quantities are typically measured by advanced meters, which tend to be interdependent or correlated. For example, active and reactive power are linked by the power factor, and unusually large deviations of this power factor over a certain period of time are an indication of anomaly. When a sequence of data is detected as abnormal, further inspection must be performed, ideally by visualization of the time series and by leveraging related metadata or unstructured data. The aim is to determine whether the anomaly is an unusual event or actually reflects bad data quality. In the second case, abnormal data are replaced by NA values which are handled during the missing data imputation stage. Alternatively, if an issue is at the system level such that a substantial part of the advanced meters are affected at the same period of time, only the portion of good quality can be retained, assuming that this is sufficient for the subsequent analysis.

The detection of abnormal time series relies on key indicators of data quality as mentioned in Section 3.3.2. The distribution of specific statistical features across the data set can be visualized in the form of histograms, and samples lying at the distribution tails might be defined as abnormal. Nevertheless, the frontier between abnormal and normal data for cleaning purposes is very subjective and is conditioned by the feature in question as well as the type and the usage of data. Regarding the percentage of missing data as illustrated in Figure 3.15, it must be considered that they are imputed at a later stage. If the percentage is too high, the resulting imputed time series might not reflect the behavior of the actual data faithfully. Hence, time series with a missing data percentage higher than a certain threshold, typically between 1% and 10%, are discarded from the data set. The exact threshold primarily depends on the tolerance level to modifications of the data properties and on the care given to the imputation stage.

When considering zero values, the acceptable percentage is mainly dependent on the measured quantity. While voltage time series should obviously not contain zero values, no general assumption can be made for active power quantities, which requires more detailed information. For example, the presence of zero values is highly unrealistic at an aggregate level (e.g., loading of a distribution transformer) but is totally normal for single devices. For smart meter measurements of residential or commercial consumers, the threshold is usually set between 5% and 20%, and further visual inspection determines if abnormal time series should be discarded. For PV production measurements, a percentage of zero values lower than 30% or higher than 70% is generally considered abnormal, and further analysis is carried out. As previously mentioned, the correlation with solar irradiance as well as the absence of substantial activity at night are good indicators for the validity of PV production time series. By experience, they should be characterized by a correlation coefficient of at least 0.6 in absolute terms. Depending on the subsequent usage of the data set, abnormal PV time series should be categorized as load profiles or simply filtered out from the data set. Conversely, actual PV production time series might be wrongly stored in a load data set, which is detectable by an abnormally high correlation with solar irradiance. Similar analysis can be performed with supposedly temperature-dependent loads based on the correlation with outside temperature, as presented in [244].

In addition, voltage measurements in the same portion of a grid must be physically correlated with each other such that a drop in the correlation coefficient with respect to neighboring voltage measurements is an indication of anomaly. Furthermore, the average or the sum of measured values in a time series can be leveraged for data cleaning purposes. Sometimes, errors in sign lead to unrealistic average values, which is fixed by replacing affected data points by their opposite value.

Finally, the comparison of the cumulative energy measured by advanced meters with the records used for billing purposes provides additional insight into possible anomalies. This is illustrated by Figure 3.18 for actual smart metered residential consumers in a one-year period. Each sample of the

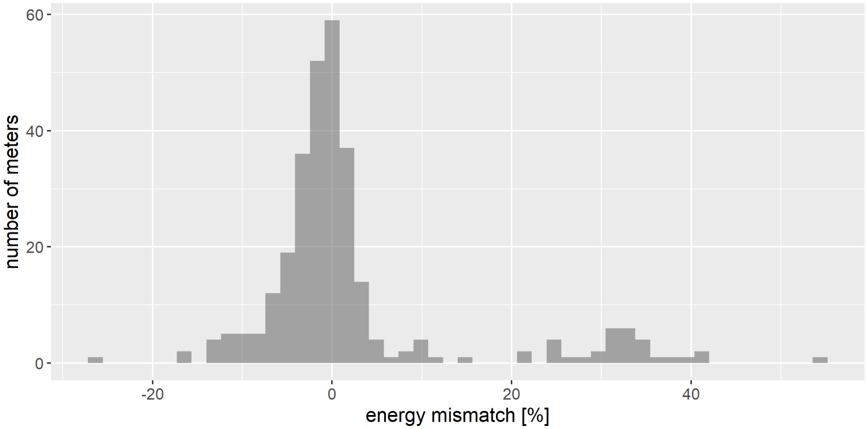


FIGURE 3.18: Relative difference between smart metered yearly energy consumption and billed energy as provided in metadata for residential consumers of a same neighborhood.

histogram corresponds to the relative mismatch in percentage between the smart metered and the billed energy:

$$\Delta E_i = 100\% \cdot \frac{E_{\mathrm{meter},i} - E_{\mathrm{bill},i}}{E_{\mathrm{bill},i}}, \tag{3.2}$$

where ΔE_i is the energy mismatch for consumer i, and $E_{\mathrm{meter},i}$ and $E_{\mathrm{bill},i}$ are the smart metered and billed yearly energy for consumer i, respectively. In general, the energy mismatch is close to zero but still varies between -27% and 53.8% for specific consumers and can even exceed 1000 kWh in absolute terms over one year. It must be noted that billing data in this example are not determined based on smart meter measurements but on the traditional electromechanical meters which have been still operated in parallel to the smart metering system. Some differences come from the fact that both systems do not systematically record the exact same load. For instance, billing data can encompass the shared consumption (e.g., heating system) in a multi-family building, whereas each smart meter records electricity exclusively used by each household. Some positive mismatches are further explained by the fact that local energy production is sometimes metered separately from energy consumption while the electricity bill considers the net energy consumption. In addition, a certain margin of error, which can typically reach 1%, is permitted for metering devices. Missing data in measurement time series also lead to a lower cumulative recorded consumption than the actual billed energy. Nevertheless, large discrepancies are considered as anomalies and can result from additional errors either in measurement data, in billing data, or even in the metadata which link measurement and billing data. If a substantial mismatch cannot be explained and fixed, the corresponding measurement time series might be discarded from the data set.

3.3.5 *Missing Data Imputation*

Missing values decrease the quality of a data set. In the context of data preparation, they must be imputed, i.e., replaced with probable values. The reason for missing data affects the choice and the design of imputation methods and must be clearly identified. Three mechanisms of missing data are commonly distinguished [245]:

- Missing Completely At Random (MCAR): The probability of missingness depends neither on other measured variables nor on the missing values. In other words, missing data are perfectly unsystematic and cannot be explained by a significant factor. For example, an accidental data transmission failure is a cause of MCAR.
- Missing At Random (MAR): The probability of missingness depends on available information but not on the missing data themselves. For example, missing data are often observed in the hour between standard time and DST.
- Missing Not At Random (MNAR): The probability of missingness directly depends on the missing data themselves. For example, the probability that the sensor of a PV panel fails can increase when the PV production increases since high sensor temperature, which is correlated to the solar irradiance, can affect its normal operation.

Based on a literature review, the authors in [246] have noticed that imputation methods perform particularly poorly in the case of MNAR since there is no proper reference example for possible values. Fortunately, missing values in AMI data sets are mostly MCAR and, to a lesser extent, MAR. In other words, missing data can typically not be explained by a known factor other than an accidental failure of the measurement sensor, data transmission system, or data management system.

For the purpose of this work, multiple imputation techniques have been implemented as detailed in the following[2]:

---

[2] For the sake of conciseness, variables are only defined once.

- Zero-order hold: Each missing value of a data gap is replaced by the last recorded value:

$$\tilde{y}_t = y_a, \tag{3.3}$$

  where $\tilde{y}_t$ is the imputed value at time step t and y_a is the value in the same time series at time step a, i.e the most recent time step before t associated to a non-NA value.

- Zero imputation: Each missing value of a data gap is replaced by a value of zero:

$$\tilde{y}_t = 0, \tag{3.4}$$

- Mean imputation: Each missing value of a data gap is replaced by the mean over all recorded values in the time series:

$$\tilde{y}_t = \frac{1}{|\mathcal{T}|} \cdot \sum_{i \in \mathcal{T}} y_i, \tag{3.5}$$

  where y_i is a non-NA value, and $\mathcal{T}$ is the set of all time steps associated with non-NA values.

- Linear interpolation: Each missing value of a data gap is replaced by linear interpolation between the most recent valid value and the next valid value:

$$\tilde{y}_t = y_{a-1} + \frac{y_{b+1} - y_{a-1}}{b-a+2} \cdot (i-a-1), \tag{3.6}$$

  where y_b is the value at time step b, i.e the next time step after t associated to a non-NA value. It must be noted that the values directly preceding y_a and following y_b are taken as reference for the linear interpolation. In fact, it has been observed that the values directly surrounding a data gap (i.e., y_a and y_b) can be corrupt[3].

- Replica from a similar day: Each missing value of a data gap is replaced by a recorded value from a similar day at the same time:

$$\tilde{y}_t = y_{t-d}, \quad \text{with } d \in \mathcal{D} = \{\pm 1\,\text{day}, \pm 2\,\text{days}, \ldots, \pm 1\,\text{week}, \ldots\}, \tag{3.7}$$

  where $\mathcal{D}$ is a set of time durations corresponding to various multiples of one day or one week. The underlying idea is that many AMI time series exhibit a one-day and/or one-week periodicity. Hence, data gaps are

---

[3] For example, in the case of a sensor failure in the middle of a time step, the corresponding energy/power measurement is often lower than it should be since it is recorded as the average observed value over this time step.

imputed by the closest complete sequence of values from another day at the same time period or from another week on the same weekday at the same time period.

- Average from similar days: Each missing value of a data gap is replaced by the average over recorded values from different other similar days at the same time:

$$\tilde{y}_t = \frac{1}{|\mathcal{C}|} \cdot \sum_{i \in \mathcal{C}} y_{t-i}, \quad \text{with } \mathcal{C} \subset \mathcal{D}, \tag{3.8}$$

  where $\mathcal{C}$ is a finite subset of $\mathcal{D}$. This is an extension of previous imputation method which considers the average sequence of non-NA values over multiple days at the same time period or over multiple weeks on the same weekday at the same time period.

- k-Nearest Neighbor (kNN): Each missing value of a data gap is replaced by the average over recorded values at the same time step from the k most similar time series in the data set:

$$\tilde{y}_t = \frac{1}{k} \cdot \sum_{l=1}^{k} y_t^{(l)} \tag{3.9}$$

  where $y_t^{(l)}$ is the value in time series (l) at time step t. The measure of similarity between two time series is based on the Root Mean Square Error (RMSE) as defined in Equation (6.18), where time steps with NA values have been discarded. Hot-deck imputation is a specific case of kNN where $k = 1$.

- Artificial Neural Network (ANN): This is a well-known ML algorithm that mimics the operation of the human brain. The *nnetar* function in the *forecast* package for R is used for imputation purposes. It trains a feed-forward ANN model with a single hidden layer and with lagged values of the time series as inputs. More information is given in [247].

- Adaptive Markov Chain Model (AMCM): This algorithm is designed based on the principle of Markov Chains combined with Gaussian Mixture Models (GMMs) with the aim of reproducing statistical properties (e.g., distribution of values) of the observed time series. A detailed description is given in Section 6.2.2.

- Status-quo: Missing data are simply preserved in the time series and will be ignored in future analysis.

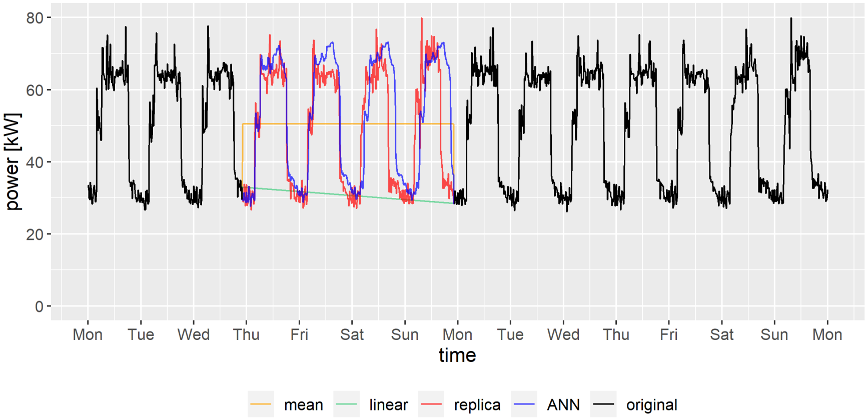


Figure 3.19: Imputation of a four-day gap in the load profile of an industrial consumer based on the mean, linear regression, replica from the following week, and ANN.

All methods presented above are frequently mentioned in the literature, with the exception of the AMCM approach as well as the replica and average from similar days. The first four techniques are quickly implemented and applicable but do not account for the variation in the data. Among all techniques, the kNN algorithm has the advantage of considering the relationship between different time series. As reviewed in [248], more advanced techniques are also suggested in the literature, which goes beyond the sole application to energy data. Especially, existing ML regression models can be adapted to tackle missing data imputation. For example, the authors in [249] propose a bi-directional missing data imputation scheme based on Long Short-Term Memory (LSTM), a special type of deep ANN. Alternatively, the authors in [250] make use of multi-objective Genetic Algorithms (GAs). However, these more advanced techniques have not been considered in this work. First, its focus is not the review and analysis of imputation techniques. The range of methods that have been implemented already offer considerable diversity for the preparation of the data sets in question. In addition, the training of advanced regression models is relatively time-consuming, which is a barrier to their application to large AMI data sets and often raises questions about the interpretability of the outcome.

Figure 3.19 illustrates the behavior of some of the imputation techniques implemented in this work on an industrial load profile with a clear one-day periodicity. Imputation by the mean and by linear regression are obviously too simplistic and fail to follow the periodic load pattern. The replica approach fills in the four-day gap with a perfect copy of the same weekdays in the following week, which seems realistic in this case. Finally, the ANN has learned the typically periodic pattern from previous days and produces a smooth but still meaningful load profile.

In this section, a comparison of the accuracy among different imputation methods is deliberately not carried out for multiple reasons. First, the imputation accuracy of most common techniques has already been evaluated in the literature, although this is mainly performed in a safe environment and usually evaluated on specific types of data. Second, the performance evaluation of imputation methods requires the artificial removal of actual values in addition to the real missing data. Since the cause of missing AMI data is hardly known, it is difficult to create artificially incomplete data which are representative of real missing data. Third, the choice of a suitable imputation method notably depends on the use that will be made of the data set. For example, the imputation of time series used in power flow simulations for the detection of LV grid congestions should reflect the actual data distribution, as discussed more concretely in Chapters 6 and 7. For the purpose of forecasting, only small data gaps of a few time steps are filled in, typically by linear regression. Larger gaps are not imputed in order not to bias the training and evaluation processes of the forecasting algorithm. Samples with remaining missing values are simply ignored by the algorithm. Fourth, the relative performance of an algorithm depends on the evaluation metric and especially on the characteristic that this metric assesses. Literature on missing data imputation techniques almost exclusively relies on point-wise metrics and rarely focuses on the time dependency among data points or on the realism of imputed data. In this regard, more details are given in Section 9.3 in the context of forecasting evaluation.

It should be reminded that time series with relatively bad quality have been filtered out from the data set at the anomaly detection stage. In this work, only time series with a limited percentage of missing values are subject to the imputation process. Different imputation methods might apply depending on the type of data. For example, voltage or PV production measurements are highly correlated with analogous measurements in the same system, which is leveraged by the kNN algorithm. Conversely, load profiles at the consumer level are relatively volatile and have little in common with each other but show similar behavior from one day to the next or on the same weekday. Therefore, the imputation of missing data with recorded values of a similar day at the same time seems reasonable.

Moreover, the function handling missing values has been designed to allow the use of various imputation methods according to the length of the data gap. The imputation of small gaps (e.g., up to one hour) largely benefits from the information of the closest available values in the time series. In this sense, linear regression and zero-order hold are considered efficient approaches. For gaps of medium size (e.g., between one hour and one day), it is judicious to consider the time dimension assuming that the data presents some periodicity. This is handled by the AMCM approach and the replica from a similar day. In the specific example of PV production profiles, missing data at night can be seamlessly replaced by zero values. Finally, special care must be given to longer data gaps (e.g., more than one day) as their imputation can partially impact the time series characteristics. The kNN algorithm can be used if similarities are observed among different time series of the same type. Alternatively, advanced imputation methods such as ANNs might be able to grasp complex data phenomena. Nevertheless, preserving missing values is sometimes a wise solution that does not further alter the data.

To sum up, an imputation method is primarily chosen according to the size of the data gap, the type of data, and its adequacy to future data usage. Secondarily, the computational cost should not be neglected. Hence, time-consuming approaches are only considered if they offer substantial benefits in terms of data quality.

3.3.6 *Data Validation and Fusion*

A data set is generally considered clean and tidy once it has been formatted and standardized, outliers and anomalies have been fixed or filtered out, and missing values have been imputed. Nevertheless, this does not guarantee that the data set is free of errors and faithfully represents the reality. At this stage, it is sometimes sensible to carry out a second statistical analysis to first quantify how the preparation process has modified the original data set. Second, this highlights some remaining data quality issues that may have been overlooked or badly handled during the preparation process. In addition, putting into perspective different data sets in the same system is a crucial step to ensure data consistency, as mentioned previously for the anomaly detection stage. Data fusion, which consists of integrating and eventually merging multiple data sources, even allows for the creation of more information than what is provided by each individual data source.

First of all, it happens that different measurement data sets in the same system are characterized by various temporal resolutions. When it comes to merging these data sets, their synchronization to a common temporal resolution is often necessary. In practice, each data set is generally converted to the lowest available temporal resolution[4]. For the purpose of this work, a function has been implemented to adapt a data set to any given temporal resolution. Figure 3.20 illustrates the conversion process of a simple time series from 10-minute to 15-minute resolution. First, the original time series is transformed into an intermediary time series with a higher resolution. Originally recorded values are simply replicated over the corresponding periods of time. In a second step, the values of the intermediary time series are averaged over the periods defined by the target resolution, which builds the new time series. Formally, the conversion process can be expressed by the following equation:

$$y_{t,\text{new}} = \frac{1}{\Delta s} \cdot \sum_{i=t_0-0.5\cdot\Delta s}^{t_0+0.5\cdot\Delta s} y_{i,\text{old}}, \quad \text{with } \Delta s = \frac{\Delta t_{\text{new}}}{\Delta t_{\text{old}}} \tag{3.10}$$

where $y_{t,\text{new}}$ is the value of the new time series (i.e., with target temporal resolution) at time step t and $y_{i,\text{old}}$ is the value of the original time series (i.e., with original temporal resolution) at time step i. Moreover, t_0 is the middle point of time step t, and Δt_{old} and Δt_{new} are the length of time steps j and t, respectively. Finally, Δs indicates the number of time steps that are considered in the averaging operation. If Δs is a non-integer rational number (e.g., $\Delta s = 15/10 = 1.5$), values at the extremity of the sequence considered in the averaging operation are weighted according to the fractional part of Δs. The conversion process as defined in Equation (3.10) is only valid for non-cumulative data (e.g., power, voltage, current, temperature). In contrast, cumulative data (e.g., energy) must initially be converted into the corresponding non-cumulative data. Furthermore, some information is inevitably lost in the conversion to lower temporal resolutions, and original time series cannot be retrieved intact when recovering the original temporal resolution. A decrease of the temporal resolution basically consists of smoothing the data. Conversely, an increase in the temporal resolution is analogous to linear regression. The impact of temporal resolution on AMI data is analyzed in

---

[4] Please note that a lower temporal resolution means a larger time step length

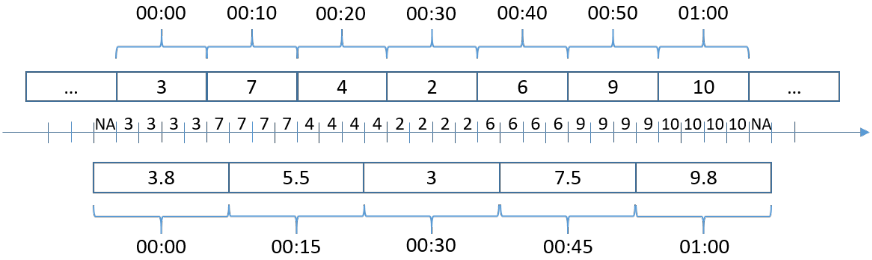


Figure 3.20: Illustration of conversion process from 10-minute to 15-minute resolution.

more detail in Chapter 5.

Moreover, in the context of distribution grids, measurement data are typically gathered at different aggregation levels (e.g., device level, customer level, transformer level) which are linked with each other. This allows for the validation of the different measurement data sets. Figure 3.21 illustrates the aggregation of active power profiles recorded by smart meters at the level of individual consumers and prosumers together with active power measurements at the transformer feeding those electricity customers. Both load profiles overlay each other and tend to exhibit similar patterns. The mismatch time series corresponds to the difference between the transformer profile and the aggregation. As expected, active power measurements at the transformer level appear to generally surpass the aggregation of smart meter measurements in the corresponding LV grid. The mismatch is nevertheless very volatile and even shows negative values for a few time steps. On the one hand, this comparison gives an indication of the remaining load and production which are not measured since the roll-out of smart meters is not complete. It must be noted that the mismatch also accounts for losses in LV lines. On the other hand, this provides further insight into potential data quality issues. There might still be undetected data errors, or the preparation process might not have been performed correctly. In addition, largely positive and negative spikes in the mismatch time series can be explained by synchronization problems among the different measurements. Metadata and their preparation are also prone to errors, which could lead to the consideration of smart meter data which do not belong to the grid in question. In any case, further investigation is required to explain unrealistic observations (e.g., negative mismatch) and increase the data quality.

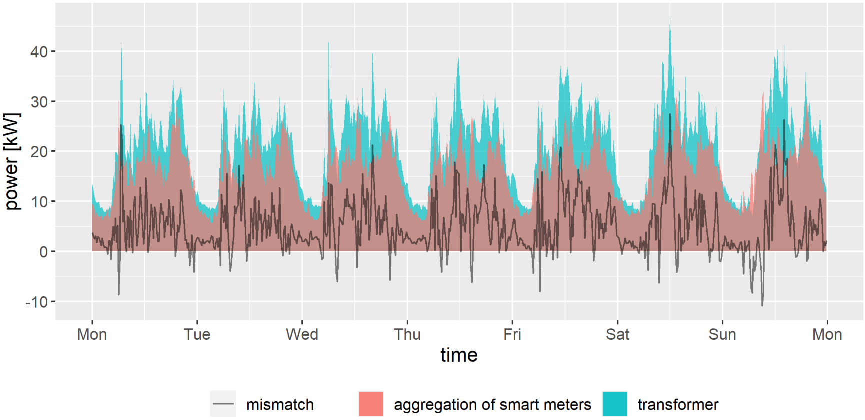


FIGURE 3.21: Weekly load profile of the aggregation of smart metered consumers in a LV grid and of the corresponding distribution transformer, together with their mismatch.

Another example of data fusion is the integration of measurement data with the grid structure, as illustrated by Figures 3.4 and 3.5. Such an integrated system can only be used for further analysis if all measurement data, grid structure data, and metadata, as well as their preparation, are of good quality. In this case, load flow simulation and state estimation help for data validation. This integration stage is particularly time-intensive due to the diversity of potential error sources and the probable need for multiple iterations in the preparation process. Among possible errors, electricity customers might be assigned to incorrect buses, line and transformer parameters might be inaccurate, grid structure and connections might be erroneous, or the binary status of circuit breakers might be wrong. Good knowledge of the different data sets, as well as good coordination with technicians responsible for data gathering, are primordial for proper data integration and validation.

3.4 CONCLUSION

To sum up, data in distribution grids are extremely diverse in many aspects. First, the roll-out of advanced meters spans from a few devices at key MV locations (e.g., substation, distribution transformers) to full smart meter penetration at end customers' premises (e.g., residential, commercial, and industrial customers). Sub-metering is rarely realized on a large scale due to its high cost, but measurement data at the appliance level are sometimes gathered in specific studies, usually for a limited period of time. It must be kept in mind that the installation of metering devices is planned in conjunction with the appropriate development of both communication and data storage infrastructures. Second, although the sampling frequency of advanced meters lies in the kHz range, their output granularity normally varies between 1 and 30 minutes, sometimes even 1 hour. This considerably influences the accuracy and quantity of recorded data. More precisely, the temporal resolution of smart meters is generally limited to 15 or 30 minutes as a trade-off between accuracy and privacy, cabinet and transformer measurements are commonly recorded every 10 to 15 minutes, and sub-metering data typically reach a one-minute resolution (or even higher for NILM purposes). Third, the type of measured quantities depends on the application. All metering devices measure three-phase sinusoidal current and voltage signals at high frequency, based on which a multitude of electrical quantities can be computed. However, their output is usually limited to a few quantities. Notably, smart meters always record active power, sometimes reactive power, and rarely voltage measurements. In addition, most smart meters currently record single-phase quantities, although power utilities start to see the need for three-phase measurements due to the unbalanced nature of LV grids. More advanced meters at the device level, but also at the cabinet or transformer levels, tend to record more quantities such as currents and voltages. Fourth, metering devices are not the only source of data in distribution grids. Metadata play an important role in providing a context for measurement data and linking different data sets in the same system. Furthermore, the network structure itself gets digitized by an increasing number of DSOs, possibly with GIS data. As traditionally performed at the transmission level, this allows for load flow simulations and state estimation also in distribution grids down to the LV level to obtain information about quantities that are not directly measured (e.g., voltages and power flows). Finally, exogenous data such as weather time series are essential to explain some behaviors observed in measurement data.

The three real-world systems presented in this chapter give an overview of the diversity and complexity of AMI data. They also highlight the limitations of synthetic data sets and test feeder models which are still largely used in the literature. Hence, this thesis intends to encourage the use of real-world or realistic data in the design of data-based approaches and in the conduction of data-based studies. This inevitably relies on close collaboration between scientists and power companies when it comes to sharing data. On the one hand, power system scientists benefit from real-world data to develop methods and carry out studies that fit reality. On the other hand, power system companies, which may not have the required internal expertise, are keener to trust scientific studies and implement some of the promising methods proposed in the literature. This naturally raises concerns regarding data privacy and confidentiality, which must be properly considered. Data pseudonymization is certainly part of the solution. Deeper discussions on privacy-preserving solutions are nevertheless out of the scope of this thesis. Besides, real-world demonstrators such as NEST, the research and innovation building of Empa (Swiss Federal Laboratories for Materials Science and Technology) in Dübendorf, Switzerland, are great initiatives to foster such collaboration between partners from the energy sector together with the scientific community [251, 252]. They allow promising technologies and operational approaches to be tested and validated in a real and open environment.

As detailed over the course of this thesis, real-world data definitely allow for the creation of realistic and meaningful case studies. However, they also come with a non-negligible drawback in terms of data quality. Due to the multiple possible sources of error between the measurement process and actual data storage, real measurement data are inevitably prone to anomalies, inconsistencies, and missing values. For their use in future applications, a preparation process is necessary to bring raw data into a formatted and standardized form in the first stage, and to detect and fix potential signs of bad data quality in the second stage. The customized data formatting and standardization stage ensures a uniform data structure and consistent conventions in terms of timestamp, missing data, language, unit, and sign. Subsequently, the standardized data preparation stage deals with outliers, anomalies, and data gaps. Proper statistical analysis and visualization also help to assess the data quality. In addition, individually prepared data sets can be compared and eventually merged to further increase the data quality but also produce additional pieces of information. For example, merging power measurement data sets from different aggregation levels allows the identification of the portion of load which is not measured yet. Such preparation process also concerns metadata and grid models which are mainly produced manually by power engineers and are to be integrated with measurement data.

Although data preparation can be standardized to a certain extent, the exact methodologies used for the different cleaning steps largely depend on the type of data and on its future application. There is clearly no predefined recipe that generally applies to all data sets, even though some guidelines are suggested in the chapter. In this context, domain knowledge and experience are certainly valuable. It must be nevertheless remembered that the preparation of data sets might slightly alter their intrinsic properties. Moreover, not all data cleaning steps must be necessarily completed at the end of the preparation process. Certain algorithms can handle data that are not totally clean and tidy. For example, most SE techniques detect and ignore bad measurement data points [253–256]. Some SE techniques are also designed to directly leverage heterogeneous data sets with different temporal resolutions [257, 258]. This avoids the loss of information when harmonizing timestamps among data sets and improves SE accuracy. In addition, some forecasting algorithms can deal with incomplete training data such that missing values imputation is not necessary anymore [259, 260]. In any case, decisions taken during the preparation process must be logged. It is important to keep track of the criteria identifying a data point as an outlier and a time series as abnormal, of the techniques used for missing data imputation, and of the samples which are discarded or modified.

Finally, the data preparation process presented in this chapter principally focuses on ensuring proper data quality. Nevertheless, clean and tidy data sets are generally further transformed to fit the needs of future applications. Often, data-based algorithms cannot directly use time series as such and require a specific features extraction stage. Features might not only come from measurement data but also from timestamps which reveal relevant temporal information and from exogenous data. More information about the actual data preparation and feature selection for each type of algorithm is given in the corresponding sections of this thesis. Data reduction techniques might also be necessary to lower the quantity and complexity of measurement data.

4

SMART METER DATA CLUSTERING AND VISUALIZATION

This chapter highlights the utility of unsupervised learning and of visualization to enhance the comprehension of large AMI data sets. They are considered big data, and their potential cannot be identified without appropriate data mining techniques. Unsupervised learning such as clustering allows for complexity reduction in order to focus on key aspects of the big data. In addition, suitable visualization of the points of interest provides power system engineers with valuable intuition in their decision-making process. As a case study, this chapter focuses on the use of the k-means clustering algorithm and leverage smart meter data from the distribution grid of the City of Basel. Its content is principally based on [261].

4.1 INTRODUCTION

The AMI of distribution grids produces a huge amount of data. End electricity customers are among the largest sources of AMI data, especially in the form of time-series measurement. Such big data potentially involve very diverse pieces of information which are difficult to understand at first glance. Data visualization is primordial for enhancing the comprehension of large data sets. As illustrated by Section 2.5, part of the work of smart grid companies leveraging AMI data consists of developing intuitive visualization tools and dashboards to obtain an at-a-glance overview. Nevertheless, the considerable volume of measurement data prevents their visualization in their original form. Beforehand, measurement data must be simplified into key features for the sake of clarity. For example, statistical indicators (e.g., average consumption, variance) can be extracted from measurement time series and displayed under the form of a histogram. This informs on the statistical distribution of customers with respect to a certain feature. Alternatively, deeper insights are offered by data mining techniques such as unsupervised learning algorithms.

An unsupervised learning algorithm is a model that is trained to discover hidden structures in a certain data set. Unlike supervised learning methods that classify instances or build a regression function, input features in unsupervised learning are not associated with a label, i.e., the outcome cannot be compared with a supposedly good answer. Unsupervised learning includes various approaches such as anomaly detection, dimensionality reduction, and clustering. Anomaly detection is part of the data preparation stage has already been discussed in Section 3.3.4. Next, Dimensionality reduction consists of compressing the original data while maintaining their most relevant characteristics [262]. Widely used in smart grid applications, Principal Component Analysis (PCA) is the most popular dimensionality reduction algorithm [263, 264]. In addition to reducing data storage overhead, it also improves data mining efficiency and is particularly helpful as a preparation step for future data-intensive applications such as forecasting [265]. Finally, the objective of clustering is to group together similar instances, also called measurements or observations in the data mining literature. The clustering outcome can be tailored to specific needs through the choice of the input features. For example, the authors in [266] intend to characterize domestic load profiles based on smart meter data. Similarly, the authors in [267] focus on the behavior of residential electricity customers for better LV network modeling and management. Alternatively, clustering of voltage measurements allows for phase identification of smart meters [268].

Multiple algorithms are proposed in the literature to perform clustering. Detailed in Section 4.2, k-means is probably the most popular clustering algorithm. It is used in this chapter for its guarantee of convergence (not necessarily to the optimum), scalability to large data sets, and ability to provide meaningful results. Nevertheless, the number of clusters must be chosen in advance, and the outcomes might depend on the initialization. A large variety of alternative clustering algorithms are also commonly leveraged for smart grid applications, especially for electrical load pattern grouping [269]. For example, hierarchical clustering is a tree-based algorithm that builds a hierarchy of clusters [270]. The expectation-maximization algorithm creates clusters based on statistical distributions. Some types of ANN (e.g., self-organizing map) are also designed for clustering purposes. It must still be noted that this chapter focuses on the manner to leverage clustering for enhanced smart meter data comprehension and does not intend to achieve a rigorous evaluation of the clustering performance. Hence, k-means could typically be replaced by an alternative algorithm for the purpose of this work.

With the help of smart meter data clustering and appropriate visualization, this chapter aims to provide useful insight into the different types of electricity consumers at a city level. As presented in Section 2.4, this provides visibility into distribution grids and can serve as a basis for DR programs. The application of clustering techniques to smart meter data for consumer profiling purposes is common in literature. Some examples are cited in the following sections. Proposed studies principally focus on the performance of different clustering algorithms on smart meter data. In contrast, the main contributions of this chapter lie in the choice of input features, in the diverse visualization of clustering outcomes, and in their application-oriented interpretation. This analysis is part of the project "Optimized Distribution Grid Operation by Utilization of Smart Metering Data" in collaboration with Adaptricity [221]. It leverages smart meter data available in the City of Basel. In this case, data are pseudonymized such that no additional information on the type or habits of individual consumers is accessible, except the load profiles. However, the location of each DCU and the assignment of smart meters to the respective DCUs are known. This gives an indication of the approximate location of each consumer across the city and is leverage for visualization purposes.

The remainder of this chapter is organized as follows. Section 4.2 provides the necessary theoretical basis concerning the widely used k-means clustering algorithm. Section 4.3 briefly describes the preparation of smart meter data used in the case study and presents different types of clustering features that can be extracted. Section 4.4 illustrates how the clustering outcome can be visualized and interpreted. Notably, temporal and spatial representations are detailed in Sections 4.4.1 and 4.4.2, respectively. Finally, Section 4.5 summarizes the ideas developed in this chapter and enhances the benefits of clustering and proper visualization for power system data analysis.

4.2 K-MEANS CLUSTERING ALGORITHM

k-means is a popular clustering algorithm that is particularly easy to apply and interpret. All popular numerical computing tools (e.g., R, MATLAB, Python) contain a library with an implementation of the k-means algorithm. Initially, features need to be extracted from a clean and tidy data set to serve as benchmarks for the formation of clusters. The number of clusters K must be defined beforehand and mostly depends on the clustering purpose, the data diversity, the number of features, and the type of features. A couple of clusters might be sufficient if the main types of loads are of interest. However, several dozen groups enable the capture of more subtle differences amongst the clusters and can reveal uncommon consumers. More details about the appropriate number of clusters and the selection of suitable features for power system analysis are provided in the following sections.

Formally, the k-means clustering algorithm aims to partition all instances in a data set into k clusters such that the Within-Cluster Sum of Squares (WCSS) is minimized [271]:

$$\text{WCSS} = \min_{\mathcal{S}} \sum_{i}^{k} \sum_{x_l \in \mathcal{S}_i} \|x_l - \mu_i\|^2, \tag{4.1}$$

where WCSS is the sum of squares of the Euclidean distances (i.e., 2-norm distances) of each instance to its respective centroid. $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k\}$ is the set of all k clusters, $x_l \in \{x_1, x_2, \ldots, x_n\}$ is one of the $n(\geqslant k)$ instances, and μ_i is the centroid (i.e., mean of instances) of cluster $\mathcal{S}_i$. Each instance is a d-dimensional vector (i.e., with d features).

Figure 4.1 illustrates the process of k-means clustering in order to build $k = 3$ clusters based on the Fisher's Iris data set [272]. This well-known data set is often used in pattern recognition and statistical classification. It consists of four features (i.e., sepal length and width, and petal length and width) of 150 flowers belonging to three iris species. In the figure, feature 1 and 2 are the sepal length and width, respectively. All features are first normalized according to the min-max feature scaling:

$$x'_f = \frac{x_f - \min(x_f)}{\max(x_f) - \min(x_f)}, \quad \forall f \in \{1, 2, \ldots, d\} \tag{4.2}$$

where $x_f = (x_{1,f}, x_{2,f}, \ldots, x_{n,f})$ and $x'_f = (x'_{1,f}, x'_{2,f}, \ldots, x'_{n,f})$ are the original and normalized vectors of feature f, respectively. In addition, $\min(x_f)$ and $\max(x_f)$ are the minimum and maximum values of feature vector x_f, respectively, and d is the number of features. Next, $k = 3$ data points called "cluster centroids" are initialized. Different initialization approaches are suggested in literature. For example, k instances in the data set can be arbitrarily picked as the first centroids. *k-means++* is another algorithm for the selection of initial values, where only one centroid is chosen totally randomly. Subsequently, the Euclidean norm between that first centroid and all remaining instances is computed, which is used to define a weighted probability distribution from which the next centroid is picked randomly. This process is repeated until all centroids are chosen. The choice of the first centroids is crucial since it influences the convergence of the method and the final cluster formation. *k-means++* usually outperforms random initialization. After the initialization phase, k-means is based on an iterative process where each iteration consists of two consecutive steps: cluster assignment and cluster update. The cluster

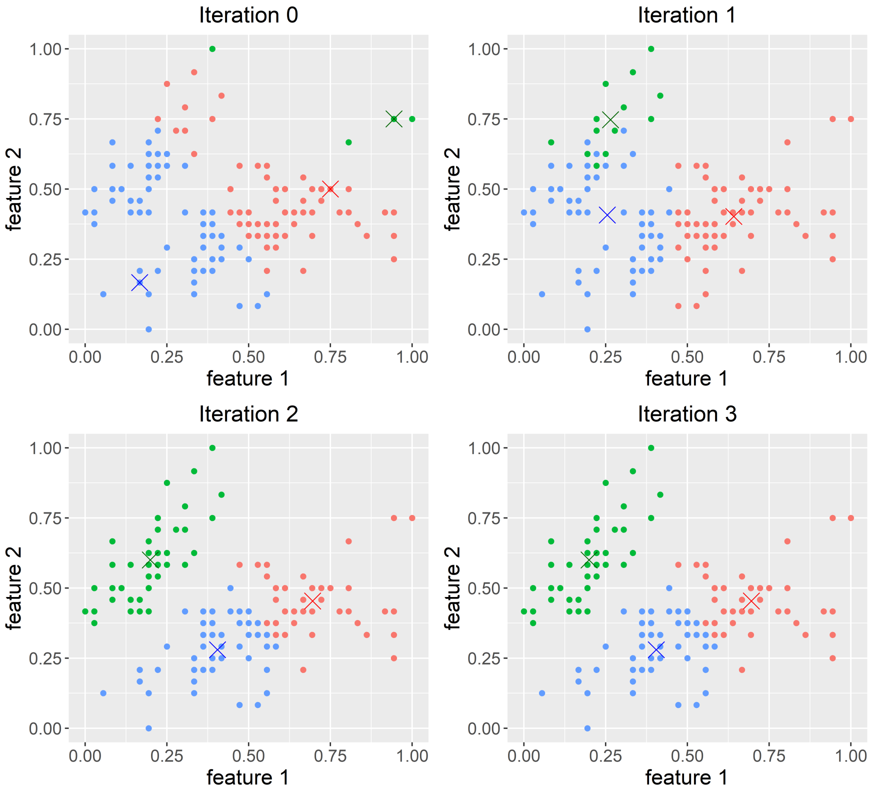


Figure 4.1: Iterative process of the k-means clustering algorithm to build three clusters out of the Fisher's Iris data set.

assignment step allocates each instance to the nearest centroid in terms of squared Euclidean distance:

$$\mathcal{S}_i^{(t)} = \left\{ x_l : \|x_l - \mu_i^{(t)}\|^2 \leqslant \|x_l - \mu_j^{(t)}\|^2, \quad \forall j \in \{1, \ldots, k\} \right\}, \tag{4.3}$$

where $\mathcal{S}_i^{(t)}$ and $\mu_i^{(t)}$ are cluster $\mathcal{S}_i$ and centroid μ_i built at iteration t, respectively[1]. The cluster update step moves each cluster centroid to the average of all respective instances:

$$\mu_i^{(t+1)} = \frac{1}{|\mathcal{S}_i^{(t)}|} \cdot \sum_{x_l \in \mathcal{S}_i^{(t)}} x_l, \tag{4.4}$$

---

[1] Remaining parameters and variables are defined together with Equation (4.1).

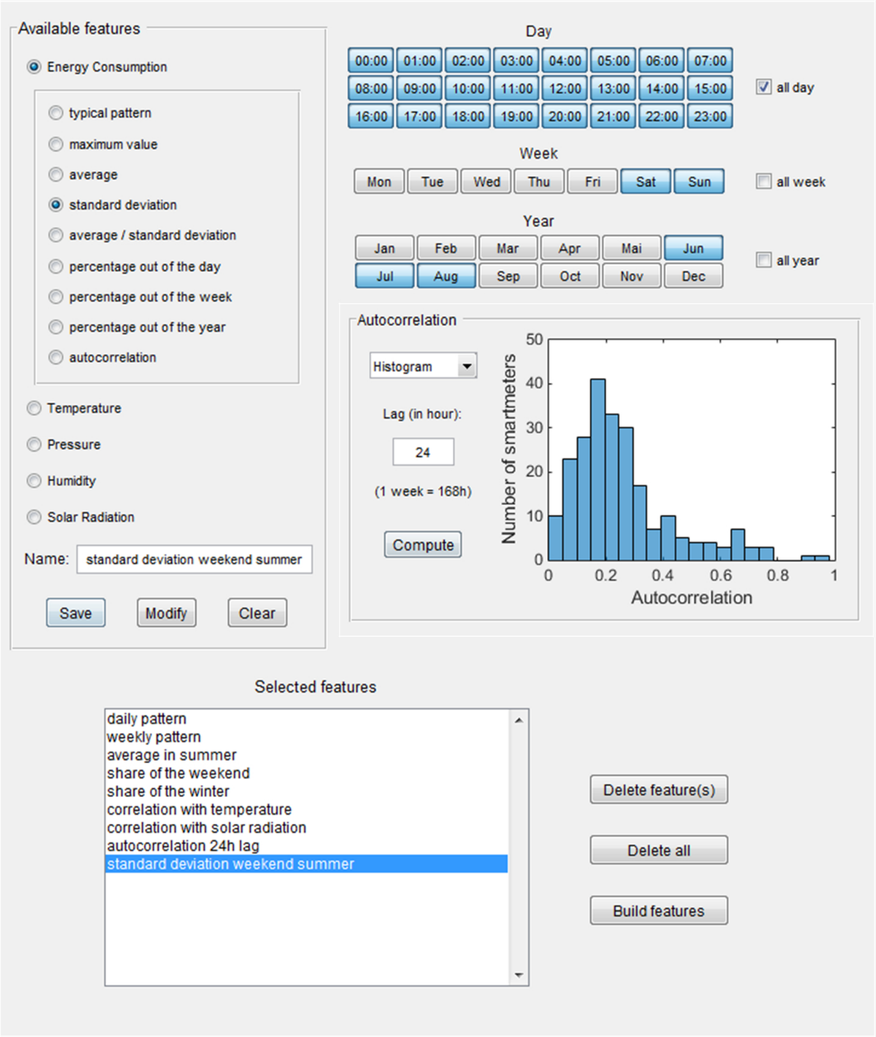
where all parameters and variables are defined previously. Both steps are iterated until all centroids stabilize, which implies that the algorithm converges. In the case of the Iris data set, three iterations of k-means are necessary for convergence.

Alternatively, the k-medoids clustering algorithm selects actual instances as cluster centroids and does not necessarily rely on the Euclidean norm as distance metric. Furthermore, k-medians clustering is a variation of k-means clustering, where each centroid corresponds to the median (instead of the mean) of all instances in a cluster. Hence, clusters are optimized based on the 1-norm distance metric (instead of the squared 2-norm distance metric).

4.3 DATA PREPARATION AND FEATURES EXTRACTION

This study leverages smart meter data from the distribution grid of the City of Basel (see Section 3.2.1 for detailed information on the data set). Good quality data are a necessary condition to obtain meaningful outcomes from a learning algorithm in terms of accuracy and interpretability. As mentioned in Section 3.3, data preparation is primordial and depends on the application. In this case, only active power profiles longer than one year and with a minimum energy consumption of 100 kWh/year are preserved. In addition, smart meter time series showing more than 10% of missing values or exhibiting data gaps larger than two weeks are discarded from the data set. In terms of missing values imputation, data gaps smaller than one hour are filled by linear interpolation. Larger data gaps are imputed by the average from a couple of surrounding weeks. Consequently, the clean and tidy data set consists of more than 30’000 time series with a 15-minute resolution and a duration going from 12 to 30 months. This data set is about 16 GB large, and its analysis is considered a Big Data problem in the context of power systems. In addition, following meteorological time series at 15-minute resolution from a weather station in the City of Basel are also considered: temperature, pressure, humidity, solar radiation.

According to the type of partitioning that is of interest among consumers, diverse features must be extracted from smart meters and weather profiles. It is particularly important to select the right amount of information that is representative of the diversity of consumers without overwhelming the clustering algorithm with unnecessary data. In any case, features represent a considerably lower amount of data with respect to the original time series. Figure 4.2 illustrates a MATLAB-based Graphical User Interface (GUI) and gives an overview of the variety of features that can be selected for



Available features

● Energy Consumption

○ typical pattern
○ maximum value
○ average
● standard deviation
○ average / standard deviation
○ percentage out of the day
○ percentage out of the week
○ percentage out of the year
○ autocorrelation

○ Temperature
○ Pressure
○ Humidity
○ Solar Radiation

Name: standard deviation weekend summer

Day: 00:00 01:00 02:00 03:00 04:00 05:00 06:00 07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00 16:00 17:00 18:00 19:00 20:00 21:00 22:00 23:00 — ☑ all day

Week: Mon Tue Wed Thu Fri Sat Sun — ☐ all week

Year: Jan Feb Mar Apr Mai Jun Jul Aug Sep Oct Nov Dec — ☐ all year

Autocorrelation: Histogram

Lag (in hour): 24

(1 week = 168h)

Selected features

daily pattern
weekly pattern
average in summer
share of the weekend
share of the winter
correlation with temperature
correlation with solar radiation
autocorrelation 24h lag
standard deviation weekend summer

FIGURE 4.2: Graphical user interface for the selection of clustering features.

the clustering analysis. Besides the usual statistical measures like the mean power consumption, the standard deviation, and the maximum power value, some more advanced features require the extraction of multiple values. For example, a typical pattern consists of computing and normalizing the average profile over a predefined period like a day or a week. Based on 15-minute resolution data, this results in the creation of 96 features for a typical daily pattern. In addition, smart meter data provide insight into daily, weekly or seasonal fluctuations for each single metered consumer. For example, the power requirements of a household, an office, or a shop generally vary over the week, especially at the weekend. In order to group consumers according to their fluctuations over the week, multiple variations of "percentage out of the week" can be selected. This feature represents the consumption share of a predefined period of the week with respect to the entire week [2]. Furthermore, the autocorrelation function, ideally with a lag of one day or one week, reveals the periodicity of power consumption profiles. Moreover, the influence of weather variables on electricity consumption is reflected in the respective correlation coefficient. The ability of a clustering algorithm to automatically point out weather-sensitive loads can contribute to the decision-making process of DSOs. Multiple features can also be combined, which gives additional value to the clustering analysis in comparison to pure statistics. Finally, extracting features from the entire measurement period may not be necessary. For grid operation purposes, a good analysis tool should give DSOs the possibility to focus on supposedly problematic hours or days. In addition, it should be able to identify customers with the largest impact on grid operation, e.g., possibly leading to voltage band violations or characterized by a high DR potential.

4.4 VISUALIZATION OF CLUSTERING OUTCOME

In any data analysis process, proper visualization is primordial to go from pure information to actual knowledge on which informed decisions can be made. This section presents and discusses the outcome of different k-means clustering processes based on two visualization approaches. On the one hand, the temporal representation visualizes clusters according to the average daily load pattern of their respective consumers. The difference between clusters still depends on the feature selection. In fact, this representation focuses on the link between a certain set of features and the temporal pattern of consumers' load. On the other hand, the spatial representation leverage information about the consumer's location. By displaying the share of each cluster in different sections of a map, this representation highlights the distribution of various types of consumers across a distribution grid. Even based on the same clustering outcome, each representation focuses on distinct aspects of the consumers' characteristics. Moreover, both temporal and spatial

---

[2] Similar reasoning applies for "percentage out of the day" and "percentage out of the year".

representations can be combined, e.g., in an interactive tool, to provide a more comprehensive overview of the diversity of consumers.

4.4.1 *Temporal representation*

The temporal representation relies on normalized typical daily profiles to visualize the different clusters. More concretely, the average daily profile is first computed for each original load profile:

$$y'_i = \frac{1}{N} \cdot \sum_{t \in \mathcal{I}} y_t, \quad \forall i \in \{1, 2, \ldots, D\}, \quad \text{with } \mathcal{I} = \{i, 2i, \ldots, Ni\},\ N = \frac{T}{D}, \tag{4.5}$$

where y'_i is the value of the average daily profile at time i and y_t is the value of the original time series at time t. In addition, $\mathcal{I}$ is the set of all time steps referring to the same time i of the day, T is the total number of time steps in the original time series, D is the number of time steps per day, and $N \in \mathbb{N}$ is the number of days in the time series. In a second phase, min-max normalization is performed as defined for feature scaling in Equation (4.2). Finally, normalized daily profiles are represented by cluster according to the clustering outcome. For the sake of clarity, the average profile in each cluster is indicated by a black curve.

In the following subsections, two different clustering processes are discussed based on their temporal representation. First, consumers are grouped according to their typical daily load pattern, which is also commonly proposed in the literature. Second, consumers are clustered according to their mean energy consumption, which highlights the correlation between the consumer size and its typical load profile.

4.4.1.1 *Typical Daily Load Pattern*

The first clustering example allows for clear partitioning of consumers according to their typical load pattern. Such clustering analysis is common in literature. For example, the authors in [266] intend to create representative electricity load profile classes for the domestic sector in Ireland based on their diurnal, intra-daily, and seasonal variations. Alternatively, the authors in [269] carry out a comprehensive performance comparison of multiple clustering algorithms based on normalized representative load patterns. In [273], representative consumption profiles are also created by clustering to serve as the basis for the classification of new electricity customers. In contrast to existing studies that principally analyze the clustering performance, this piece

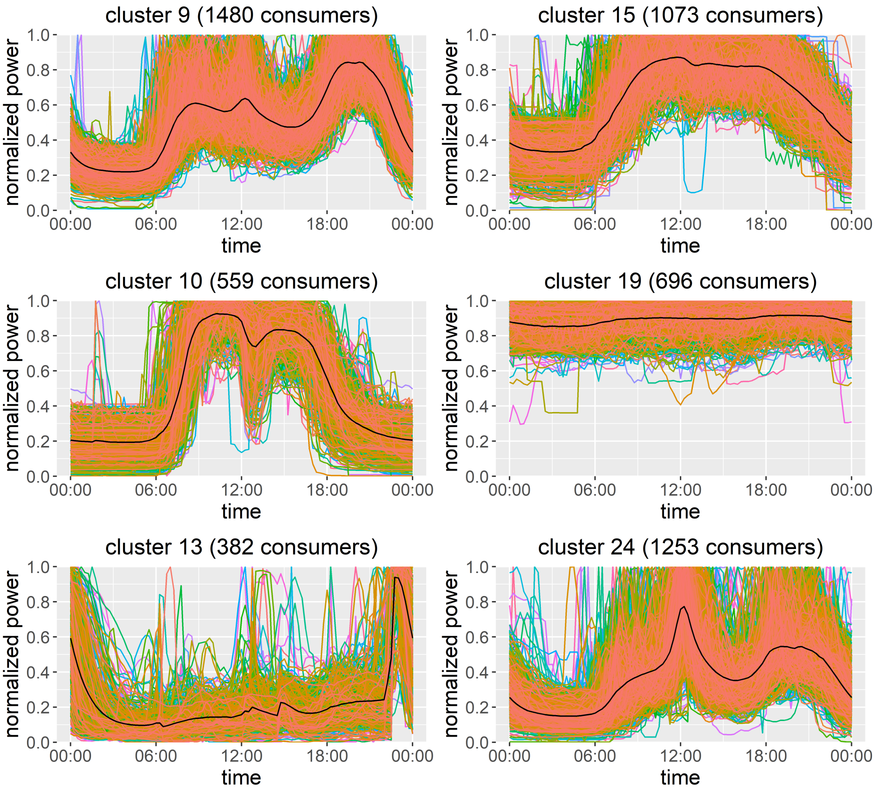


Figure 4.3: Clustering of normalized typical daily load profiles.

of work focuses on the entire clustering process, including the interpretation of the clustering outcome. Although beyond the scope of this study, it is probable that the use of alternative clustering algorithms leads to a different partitioning.

The proposed case study directly uses the normalized typical daily profile of each consumer as a feature. Based on smart meter data with 15-minute granularity, 96 features are extracted from each consumer's load profile before being supplied to the clustering algorithm. In this specific case, visualized data points directly correspond to the features, and the black curve represents the cluster centroids. As a rule of thumb, the number of clusters is fixed to 25 in order to better detect unusual loads while having a sufficient amount of consumers per cluster.

The clustering outcome is illustrated by Figure 4.3, where the analysis is restricted to the six most distinctive clusters. In this example, cluster 9 contains the largest number of consumers. It likely consists of households as suggested by their average profile. Indeed, an increase in demand appears from 6:00 on and at noon, indicating power-consuming activities like cooking. Nevertheless, most of the demand occurs after working hours. Inhabitants probably come back from work, turn on the lights, use cooking devices and the dishwasher, maybe do laundry, and watch television. It must be noted that a majority of clusters that are not shown here also exhibit similar profiles. This suggests that most smart meters are installed at residential customers' premises. Furthermore, offices seem to be grouped in cluster 10, considering the characteristic drop at lunchtime and a very low consumption outside of regular business hours. Cluster 13 mostly consists of ripple-controlled units such as boilers that are programmed to start working based on the time of the day. This is particularly clear at 21:00, where the electricity price switches from high tariff to low tariff regime in Switzerland. In cluster 15, power is mainly consumed between 8:00 and 20:00, which is indicative of the needs of a shop or a department store where load variations during opening hours also depend, to a limited degree, on the number of visitors. In contrast, loads in cluster 19 have relatively constant power consumption on average. This can be typical of 24-hour active industrial loads. Nevertheless, it must be kept in mind that profiles represent an average over a large number of days and do not reflect the behavior on a single day. For example, the typical daily pattern of thermostatically controlled loads appears relatively constant due to the averaging effect, even if their instantaneous status is either ON or OFF. Such representation still gives insight into the daily profile of an aggregation of multiple similar loads. Finally, load profiles in cluster 24 are characterized by a significant peak during lunchtime and relatively higher values in the evening (i.e., outside regular working hours). This cluster probably consists of restaurants, cafeterias, and maybe a few households. Such analysis can obviously be extended to an entire week, where a considerable power drop shall be visible at the weekend for non-residential categories.

4.4.1.2 *Mean Energy Consumption*

In contrast to the clustering analysis commonly proposed in the literature, features of smart meter data must not necessarily be related to the typical load pattern. For example, the authors in [274] innovate by making use of the variance and of the responsibility factor at different peak hours to identify customers having a higher peak share. Further interesting analysis is the

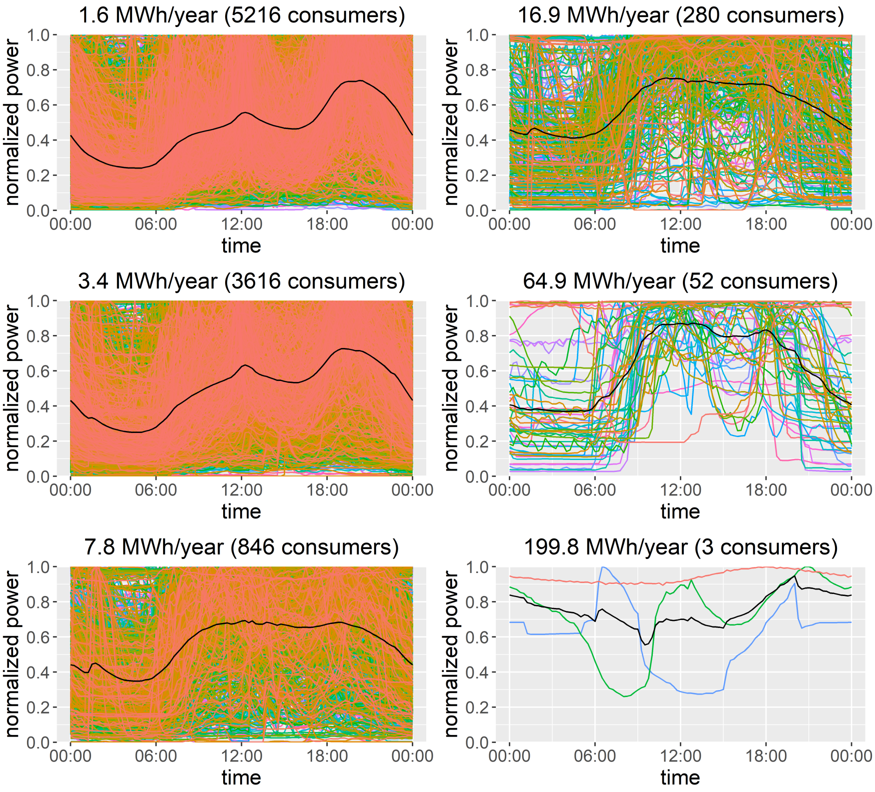


FIGURE 4.4: Typical daily profiles of smart metered loads clustered according to their mean energy consumption.

evolution of the typical load profile according to the mean energy consumption, as shown in Figure 4.4. In this case, there is only one feature, the mean energy consumption. In addition, fifteen clusters are created, but only six clusters are visualized for the sake of clarity. On average, small consumers exhibit a typical household pattern even if their individual profile can be of any shape. Nevertheless, the more energy is consumed, the more rectangular the average load profile becomes, while the usual household evening peak tends to vanish. This is explained by the higher share of commercial and industrial loads that are mainly active during regular working hours with fairly constant energy demand. Although not displayed, this observation is confirmed when considering the typical weekly profile. Small consumers show a higher activity at the weekend, which gradually decreases with rising mean energy consumption. In addition, the size of clusters drastically decreases when the mean energy consumption increases, which confirms the larger share of residential loads in the data set. Finally, the last cluster indicates that no specific trend can be assumed for very large consumers.

Even if the clustering process relies on only one feature, it allows for partitioning. This is an advantage over simple statistical analysis such as one-dimensional histograms. In this sense, similarities between consumers in a certain cluster can be inferred. Hence, a DSO can potentially get insight into the characteristics of a customer on the sole basis of its electricity bill. Furthermore, such unsupervised learning can easily be combined with other features or focus on a smaller set of customers, e.g., living in the same neighborhood. It must nevertheless be kept in mind that clustering basically distinguishes main categories over a large set of consumers. It does not replace a more thorough and specific analysis at the consumer level.

4.4.2 *Spatial Representation*

In addition to the temporal dimension, smart meter data are also characterized by a location. This spatial information can be integrated into the data mining process for better big data comprehension. Hence, this section presents an interactive Leaflet-based visualization tool that combines clustering outcomes with spatial information [275]. This visualization tool implemented in JavaScript has been designed to allow IWB (i.e., DSO of the City of Basel) to access data mining knowledge in a more intuitive way. More concretely, clustering outcomes are displayed in the form of pie charts on the city map based on the location of smart meters. In the case of the City of Basel, the exact location of all smart meters has not been made available for privacy-preserving reasons. However, the location of DCUs and the list of corresponding smart meters are known, which is sufficiently precise to obtain a good overview of the variety of smart metered consumers spread over the city. It must be reminded that the AMI of the City of Basel consisted of 387 DCUs at the time of data preparation, with an average of 70 metering devices per DCU. While DCUs represent the highest spatial resolution, the interactive visualization tool still allows for a broader overview depending on the zoom level. Hence, one pie chart possibly reflects information from multiple DCUs in the same neighborhood at lower spatial resolution. Moreover, the visualization tool is designed to display, among others, typical daily load profiles by clicking on specific customers. In this sense, both temporal and spatial dimensions of smart meter data are visible in the same environment, based on which k-means clustering adds a layer of understanding.

Depending on the focus of interest, different clustering features must be selected. The following subsections detail how spatial representation enhances the interpretation of two clustering examples. First, smart metered customers are clustered according to their share of consumption at different periods of the day. The second example considers the correlation of load profiles with the outside temperature.

4.4.2.1 *Intra-Daily Consumption Share*

Figure 4.5 spatially represents smart metered loads clustered by their share of energy consumption at different periods of the day. As suggested by [276], a day is divided into five representative time periods, i.e., early morning (7:00 to 9:00), morning (9:00 to 13:00), afternoon (13:00 to 17:00), evening (17:00 to 21:00) and night (21:00 to 7:00). Hence, each clustering feature represents the percentage of energy used in each representative time period:

$$\text{share}_p = 100\% \cdot \frac{\sum_{j \in \Omega_p} y_j}{\sum_{i=1}^{T} y_i}, \tag{4.6}$$

where share_p is the feature associated with time period p, Ω_p is the set of all time steps within time period p, and T is the total number of time steps in the time series. Per definition, all features sum up to 100% for each consumer. This clustering process is analogous to the clustering example presented in Section 4.4.1.1 at the difference that each feature does not refer to a specific time instant but a specific time period in the day. For the sake of clarity in the visualization tool, only five clusters are created with the k-means algorithm. The box at the bottom right corner of Figure 4.5 summarizes the cluster characteristics. More precisely, it includes the number of consumers per cluster as well as the values of cluster centroids (i.e., mean feature values). In this case, cluster 1 (red) is the smallest group with less than 5% of all smart metered consumers and consists of “night owls” who mainly consume overnight (i.e., 52.3% energy consumption between 21:00 and 7:00). Most loads in cluster 1 are located in areas with a majority of apartment buildings, like in the most easterly neighborhood of the City of Basel. Their typical daily profile reveals that they are especially active from 22:00 on (i.e., off-peak tariff). This gives a good indication of the buildings equipped with electric boilers, which implies higher power flows at night in these areas. Since these loads are price-sensitive, they are good candidates for DR programs. Conversely, consumers in cluster 3 (yellow) are principally active during

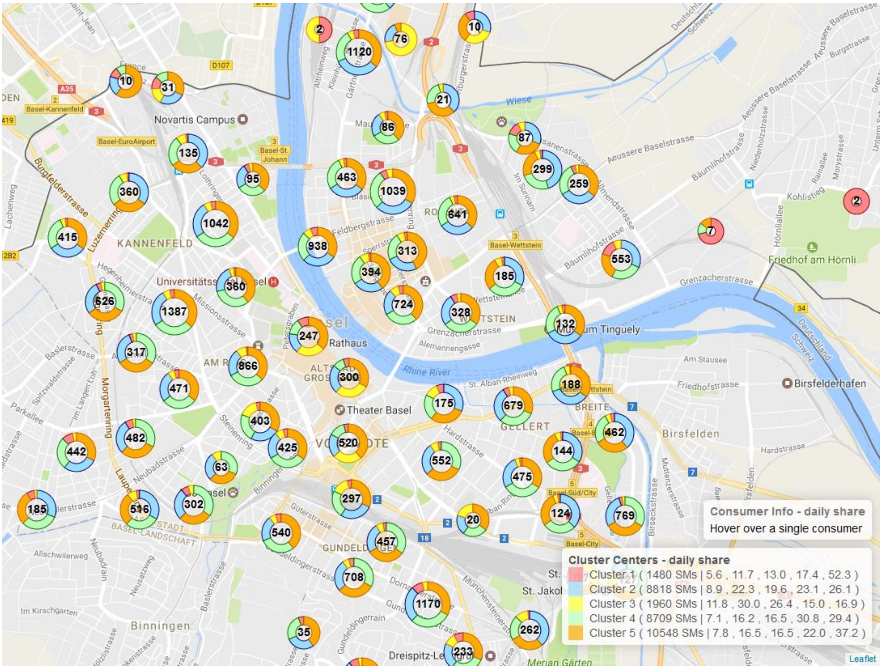

Novartis Campus
KANNENFELD
Rhine River
Theater Basel
Rathaus
Birsfelden
Friedhof am Hörnli
Birsfelderhafen
Binningen
Consumer Info - daily share
Hover over a single consumer
Cluster Centers - daily share
🟥 Cluster 1 (1480 SMs | 5.6 , 11.7 , 13.0 , 17.4 , 52.3)
🟦 Cluster 2 (8818 SMs | 8.9 , 22.3 , 19.6 , 23.1 , 26.1)
🟨 Cluster 3 (1960 SMs | 11.8 , 30.0 , 26.4 , 15.0 , 16.9)
🟩 Cluster 4 (8709 SMs | 7.1 , 16.2 , 16.5 , 30.8 , 29.4)
🟧 Cluster 5 (10548 SMs | 7.8 , 16.5 , 16.5 , 22.0 , 37.2)
Leaflet

FIGURE 4.5: Spatial representation of smart metered loads clustered according to their share of energy consumed at different periods of the day (early morning, morning, afternoon, evening, night).

business hours and are mainly concentrated in the old city center where shops, museums, offices, and restaurants are located. They account for 6% of the total number of customers. The remaining three clusters encompass the large majority of IWB's customers. They are well spread across the entire distribution grid. Cluster 2 (blue) is rather active between 9:00 and 19:00 and might include offices, restaurants, and a few households. Clusters 4 (green) and 5 (orange) seem to contain a considerable number of residential loads and probably restaurants since the corresponding consumers are characterized by relatively high activity in the evening.

To sum up, the interactive tool facilitates the visualization of energy requirements in different parts of the grid and at different periods of a typical day. A relatively high and well-distributed smart meter penetration is nevertheless required for a representative overview of the actual situation. By extension, this concept can be adapted to longer periods, such as a whole year, to focus on local seasonal variations.

4.4.2.2 *Correlation with Temperature*

Figure 4.6 exhibits the exact same loads as in Figure 4.5, but clustered according to their correlation with the outside temperature. First, it must be noticed that Swiss households are usually not equipped with air conditioners and electrical heating systems are less common than gas-fired heating systems. Nevertheless, IWB launched a campaign to gradually replace traditional heating systems with electrical heating systems. By analyzing clustering outcomes, it appears that half of the load profiles are barely influenced by the temperature. They belong to cluster 1 (red), with an average correlation coefficient of -0.02. They are most probably not equipped with an electrical heating system. Cluster 3 (yellow) is the second largest group with 30% of all customers. The load profile of its members is also negligibly correlated with the temperature. Their slightly positive average correlation coefficient of 0.06 can simply be explained by higher consumption during the daytime, i.e., when temperatures are naturally higher. Moreover, cluster 4 (green) is relatively small, but corresponding loads have the largest positive correlation coefficient (i.e., 0.27 on average). Even if these consumers have a higher electricity demand during the daytime, they seem to be influenced by warm weather conditions. Indeed, they are almost exclusively located in shopping areas or at the main football stadium, where air conditioners are usually running on hot days. Furthermore, negatively correlated consumers are divided into two groups. First, cluster 5 (orange) consists of many customers who tend to consume slightly more with decreasing temperature (i.e., average correlation coefficient of -0.12), which might indicate the presence of electrical heating systems. Almost all neighborhoods of Basel contain a small share of these consumers to varying degrees. Second, cluster 2 (blue) shows the highest negative correlation with temperature. On the one hand, it consists of loads that are naturally very temperature-sensitive. On the other hand, price-sensitive loads like boilers are also part of this cluster. In addition to a potential impact of low temperatures, their mainly overnight consumption contributes to the negative correlation.

Although most of the electricity customers are mainly active during periods of the day with naturally higher temperatures, a majority of them are still negatively correlated with the temperature. This suggests that this incidental correlation effect is limited. Based on this clustering analysis, DSOs can first gain insight into the grid areas that require a relatively higher electricity supply during extreme weather conditions. In addition, clusters with higher absolute correlation coefficients can be categorized as good candidates for DR programs. This nevertheless assumes that temperature-sensitive consumers

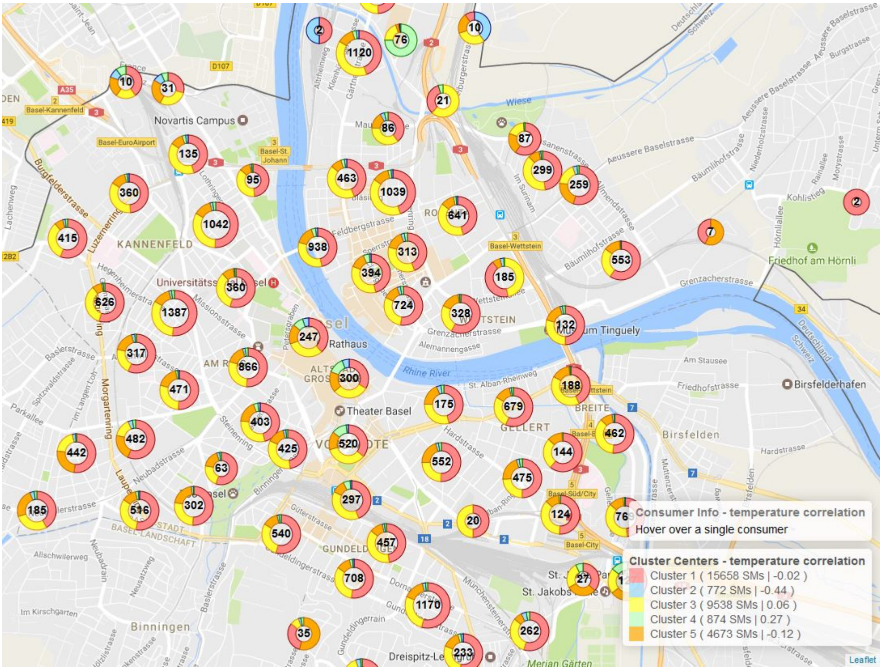

Consumer Info - temperature correlation
Hover over a single consumer
Cluster Centers - temperature correlation
🟥 Cluster 1 (15658 SMs | -0.02)
🟦 Cluster 2 (772 SMs | -0.44)
🟨 Cluster 3 (9538 SMs | 0.06)
🟩 Cluster 4 (874 SMs | 0.27)
🟧 Cluster 5 (4673 SMs | -0.12)
Novartis Campus
Rhine River
Theater Basel
Birsfelden
Binningen
Leaflet

FIGURE 4.6: Spatial representation of smart metered loads clustered according to their correlation with the temperature.

are equipped with boilers or HVAC systems that can offer flexibility thanks to their thermal inertia.

4.5 CONCLUSION

To summarize, the high spatial and temporal resolution provided by smart meters enables a previously unattainable degree of detail in distribution grids. Though, suitable methods are needed to lower the complexity of a large quantity of data and convert it into actual knowledge easily interpretable by power systems companies, which can eventually be integrated into decision-making support tools. Unsupervised learning techniques, and especially clustering, add value to smart meter data by grouping and putting into perspective the multiple consumers. There is a large set of clustering algorithms with various performance capabilities proposed in the literature. However, this analysis is limited to the popular k-means algorithm and principally focuses on the overall clustering process. Based on more than 30'000 smart metered consumers in the City of Basel, this chapter illustrates various clustering processes, going from proper data preparation and features extraction to the representation and interpretation of results. An interactive visualization tool has also been designed for presenting clustering outcomes in an intuitive form.

The extraction of features is the key element of a successful clustering analysis since they define the points of similarity between power consumers in order to build clusters. The variety of features basically depends on the points of interest, e.g., standard statistical metrics, a combination of such metrics, the correlation between load profiles and weather variables, and typical load profiles. Based on the k-means algorithm, the number of clusters must be chosen in advance, which is often a trade-off between interpretability and the desired level of detail. Clustering outcomes can be visualized in different ways, taking advantage of both the high temporal and spatial resolutions of smart meter data. For example, typical load profiles focus on the behavior over time of different categories of consumers, whereas the location of DCUs can be leveraged to visualize their spatial distribution on a city map. While preserving the anonymity of individual consumers, clustering of smart meter data represented at the level of DCUs provides a more comprehensive picture of the system than aggregate measurements, e.g., at the transformer level.

Apart from the data preparation phase, clustering is not a time-intensive method and only requires a small amount of data in contrast to other learning algorithms such as forecasting. In addition, such unsupervised learning does not require data labeling which is often time-consuming. More generally, useful knowledge can be gained from smart meter data without any further information concerning the type of consumers and their habits. Clustering analysis is of interest to identify the main types of consumers but also detect more uncommon loads whose behavior might considerably differ from the majority. As illustrated in the visualization tool, uncommon loads are located in specific areas of the city. Good knowledge of the different types of loads and of their share can help the DSO to cope with critical states of the distribution grid rapidly. For example, a large concentration of temperature-sensitive consumers in a certain neighborhood can heavily load the grid components in case of extreme weather conditions. The extent of such an issue is not necessarily visible on an aggregate level and can go undetected if suitable tools for analyzing and visualizing smart meter data are not available. Furthermore, the customer segmentation obtained by clustering can set the basis for the implementation of dynamic pricing and DR programs. More specific analysis is nevertheless necessary to estimate the actual flexibility potential. In any case, simple unsupervised methods such as clustering already offer a good overview of smart meter data for a wide range of applications.

5

TEMPORAL RESOLUTION AND SPATIAL AGGREGATION

This chapter discusses how the characteristics of load profiles in distribution grids are altered with respect to the temporal resolution and spatial aggregation. Based on different Costa Rican data sets, time series visualizations first provide qualitative insights, whereas standard statistical metrics allow for quantitative evaluation. Although rarely properly considered in data-based studies, it appears that the effect of both temporal averaging and spatial aggregation on load profiles is substantial, especially at low aggregation levels. The content of this chapter is principally inspired by [277].

5.1 INTRODUCTION

As presented in Chapter 4, both temporal and spatial dimensions of advanced metering devices contain valuable information. The level of detail is nevertheless linked to the data resolution. Depending on the application, measurements at different temporal resolutions and different aggregation levels are required. For example, NILM algorithms traditionally work with high-resolution measurements, i.e., with granularity in the range of seconds (or higher) and at the end-user level. In contrast, long-term load forecasting (e.g., for planning purposes) focuses on the main tendency such that measurements at an aggregate level for typical days are sufficient. In addition to the desired level of detail, the resolution of measurement data has implications for data communication and storage requirements. The higher the resolution, the larger the bandwidth of the communication system and the storage capacity. Finally, high-resolution measurement data raise privacy concerns as discussed in Section 2.4.6.

The output temporal resolution of advanced metering devices is a customizable parameter that results from a trade-off between different interests and constraints. Higher temporal granularity potentially implies more precise information but also leads to a larger amount of data and to additional risks in terms of customer privacy. In this regard, each country or even each DSO relies on its standards. For example, the European Union recommends a granularity of 15 minutes for storing historical smart meter consumption data [278]. Most EU states (e.g., Austria, Belgium, Germany, Greece, Hungary, Italy, Luxembourg, Malta, Slovenia) follow this recommendation. Higher resolutions (i.e., 5 minutes to 1 minute) are sometimes allowed for pilot projects and specific categories of customers, e.g., industrial consumers. In contrast, Ireland is equipped with 30-minute resolution smart meters, whereas Croatia, Latvia, Lithuania, Portugal, and Spain rely on 1-hour resolution smart meters. Slovakia is an exception with a granularity of 3 minutes for smart meter data. In addition, the examples of AMI data sets presented in Section 3.2 exhibit a temporal resolution of 10 to 15 minutes for cabinet and transformer measurements and one minute for sub-metering data, which is common in distribution grids.

The literature on the actual impact of temporal resolution in AMI data is scarce and focuses on specific aspects. For example, the authors in [279] study the effect of temporal averaging on the load profile of 8 houses recorded at a one-minute resolution. The authors in [280] investigate the impact of temporal resolution on the performance of multiple clustering algorithms for residential load profiles. Besides, temporal data granularity also influences smart meter privacy. Notably, the authors in [189] have performed a comprehensive sensitivity analysis on the possible detection of home appliances. They conclude that an appliance is visible in smart meter data as long as the time interval does not exceed half of its typical on-duration. The temporal resolution also impacts the recorded output power of PV systems. For example, the authors in [281] analyze how this influences the capacity configuration of energy storage systems. In addition, the authors in [282] show that the accuracy of spatio-temporal solar forecasting depends on the temporal data resolution.

Furthermore, measurement data can be recorded at different aggregation levels in distribution grids, going from single electrical appliances to the substation level. In this context, spatial resolution refers to the number of different measurement points at a certain aggregation level. Measurements at low aggregation levels (e.g., appliance, electricity user) are generally characterized by a higher spatial resolution than measurements at higher aggregation levels (e.g., transformer, substation). More concretely, a local substation supplies several dozens or of distribution transformers, each distribution transformer supplies hundreds of electricity users, and each user utilizes several dozens of electrical appliances. At the exception of smart-meter roll-out statistics as presented in Section 2.3, comprehensive information on the number of installed sub-metering devices and advanced meters at distribution cabinets and transformers is rare.

As reviewed by [283], spatial aggregation is often leveraged for the design of privacy-preserving schemes. For example, the authors in [284] and [285] study the theoretical computation and communication overheads of their proposed privacy-preserving schemes as a function of the aggregation level. However, the influence of spatial aggregation level on the achieved privacy itself is not quantified. Spatial aggregation is also relevant in load forecasting. As noticed in [286], the aggregation level can substantially influence the prediction performance. The authors in [123] have derived an empirical scaling law that describes load forecasting accuracy at varying levels of aggregation. Ensemble forecasting also benefits from a judicious grouping of consumers [131]. Nevertheless, to the best of the author's knowledge, there is no literature on the characteristics of AMI measurements themselves with respect to spatial aggregation. In addition, temporal and spatial dimensions are rarely considered together.

Accordingly, this chapter studies the influence of both temporal and spatial dimensions of AMI data on their inherent characteristics. For that purpose, the analysis leverages both available Costa Rican data sets whose preparation is presented in Section 5.2. Next, Section 5.3 studies the impact of temporal resolution on the load profile of domestic appliances and residential users. The impact of temporal resolution and spatial aggregation is then analyzed jointly in Section 5.4. Finally, Section 5.5 summarizes the primary outcomes of the study and points out the implications for future data-based applications.

5.2 DATA PREPARATION

Generally, raw data must be prepared in a way that meets the requirements of the subsequent data-based analysis without substantial bias of their actual properties. For this study's purpose, the data preparation process must preserve the statistical properties while yielding complete time series. The case studies are based on both Costa Rican data sets presented in Section 3.2. Sub-metering data are valuable for their high temporal and spatial resolutions, whereas the set of smart meter data contains a large number of examples. Their specific preparation is detailed in the following.

5.2.1 *Costa Rican Sub-Metering Data*

As presented in Section 3.2.3, the Costa Rican sub-metering data set consists of active power load profiles of residential appliances from 70 households recorded over about one week at a one-minute resolution. Due to the relatively low amount of data, visual inspection appears to be the most reasonable approach to filter out inadequate measurement time series. Load profiles of individual appliances can be very diverse such that their authenticity would be challenging to assess algorithmically. In this context, data filtering is generally based on domain knowledge and notably leverages different examples of usual appliance load profiles illustrated in [287]. In the Costa Rican sub-metering data set, the main reason for filtering out recorded load profiles is their unrealistic shape or power consumption level compared to the type of device indicated in the metadata. Hence, examples are principally discarded because of data labeling errors. In addition, the metadata indicate that the refrigerator of some households is metered together with other appliances in the same channel. In several cases, visual inspection nevertheless shows that the corresponding load profile still corresponds to the sole refrigerator, increasing the set of valid load profiles. In terms of cleaning, time series with a share of missing values higher than 5% are discarded. Subsequently, data gaps smaller than five minutes are filled by linear interpolation, and larger gaps are imputed by values from a similar day in order to preserve statistical properties. After data preparation, active power profiles of 24 water heaters, 14 refrigerators, 16 dryers, 37 kitchen appliances, and 59 lighting devices are available for the study. The remaining measurement time series are aggregated by household and labeled as “others”. Moreover, the difference between the aggregation of all individual measurements and the household’s main load is labeled as “not measured”.

5.2.2 *Costa Rican Smart Meter Data*

As detailed in Section 3.2.2, the smart meter data gathered by CNFL, one of Costa Rican DSOs, consists of several thousands of active and reactive power profiles at the end-user level. They cover a large range of residential and commercial consumers. A period of four months is selected for this study. In contrast to the sub-metering data, the measurement time interval of the smart meter data is only 15 minutes. In terms of cleaning, time series with a share of missing values higher than 5% are discarded. In addition, data gaps smaller than one hour are filled by linear interpolation. The values from a neighboring week replace the remaining data gaps. Eventually, the data preparation phase leads to clean active and reactive power profiles for 7’349 consumers.

5.3 EFFECT OF TEMPORAL RESOLUTION

This section investigates the effect of the output temporal resolution of advanced meters on the shape and statistical properties of load profiles. When recording non-cumulative data (e.g., power, voltage, current, temperature) at lower temporal resolutions than the original measurement frequency, so-called temporal averaging occurs. This means that each recorded value consists of the average over several initial measurement time steps. For the purpose of this study, original load profiles are adapted to lower temporal resolutions according to the conversion process expressed by Equation (3.10). This reflects how measurement data would be recorded at different output resolutions. In the following, the effect of temporal averaging is assessed on the profile of loads at different aggregation levels. Namely, the study focuses on single electrical appliances, individual residential users, and the aggregation of residential users. The analysis is illustrated by data from the Costa Rican sub-metering study characterized by an original temporal resolution of one minute.

5.3.1 *Individual Electrical Appliances*

The shape of load profiles highly varies among different electrical appliances. To get an idea of their shape and of the impact of temporal resolution, Figure 5.1 visualizes the load profile of three representative domestic appliances at different common temporal resolutions. The first load profile belongs to a water heater. This TCL consumes electric power to heat and maintain the temperature in the water tank within certain limits. It is characterized by a fixed rated power of 10 kW and a relatively long thermal inertia. Hence, power events are relatively short. When decreasing the temporal resolution, shorter power events are not visible anymore at the rated power due to the temporal averaging effect. From a temporal granularity of 15 minutes on, the water heater activity only appears over one or two time steps since the time interval is in the same range as the average ON duration. From this point on, the shape of the load profile is not anymore characteristic of a water heater. The second example corresponds to the load profile of a tumble dryer. This device is used occasionally, and its cycle typically lasts more than one hour. During

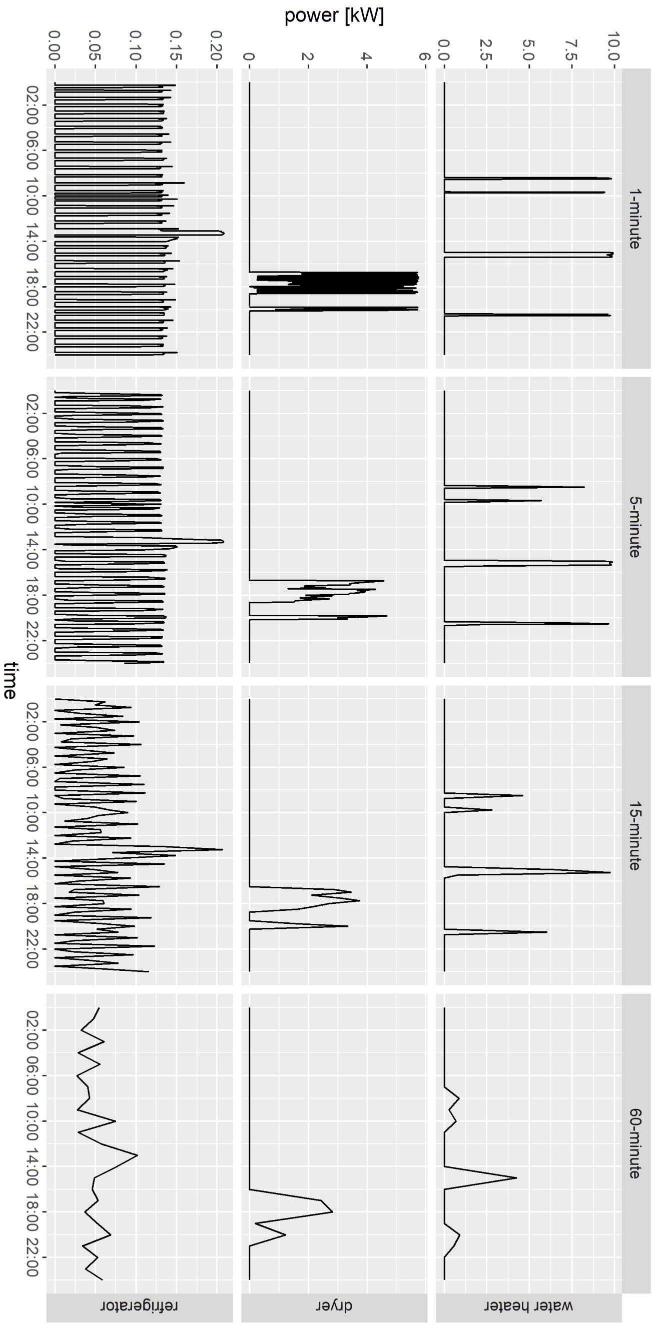


FIGURE 5.1: Active power load profiles over one day of a water heater, a tumble dryer, and a refrigerator at different temporal resolutions.

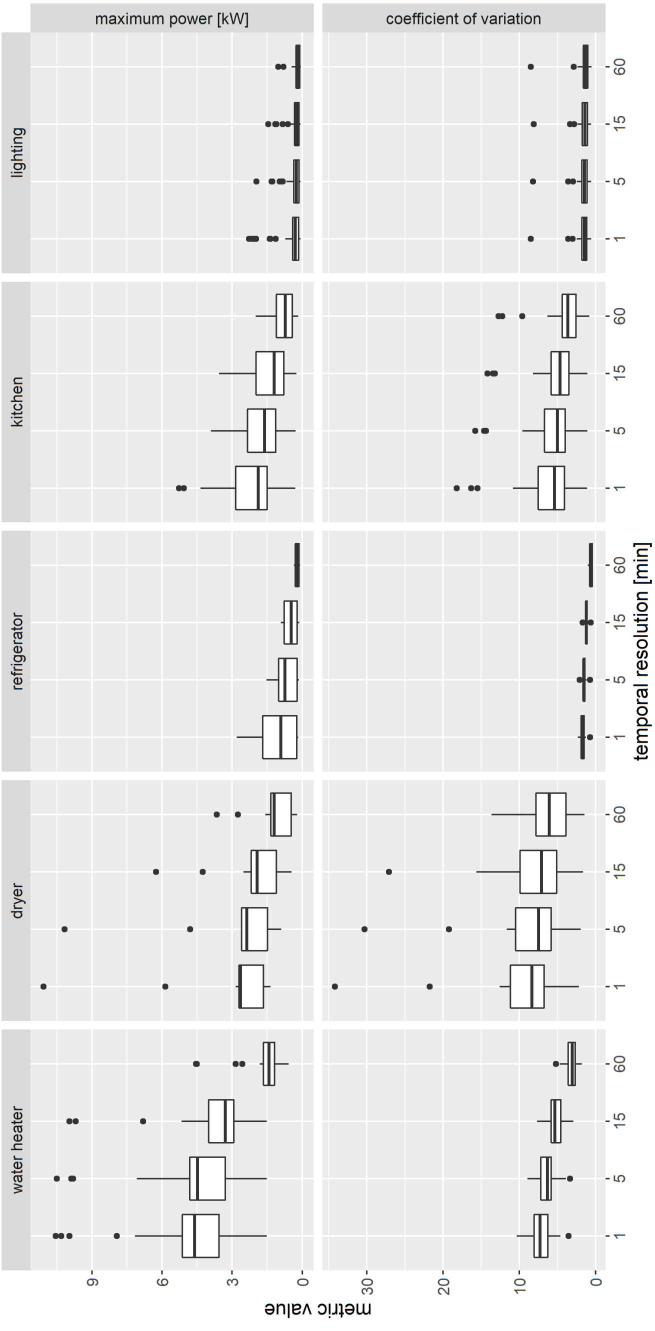


FIGURE 5.2: Box and whisker plots of the maximum value and of the coefficient of variation in active power measurements of multiple domestic appliances at different temporal resolutions.

one cycle, the power consumption varies significantly between 0 and 6 kW. Because of the frequent changes in power consumption, its typical load shape is already not recognizable from a temporal resolution of 5 minutes. The dryer load profile becomes slightly smoother with further decreased temporal granularity. Finally, the load profile of a typical refrigerator is displayed. Its rated power lies around 130 W, where a sharp power consumption peak characterizes the beginning of each cycle. In addition, a defrost cycle with 210 W power consumption appears around midday. Similar to the water heater, the refrigerator is a TCL with a typical ON-OFF consumption behavior. Nevertheless, it has to maintain the cool air within much narrower limits, and its thermal inertia is considerably lower. Hence, its duty cycle (i.e., the fraction of a cycle in which it is ON) is only slightly lower than 50%. Since this refrigerator has a mean ON duration of 20 minutes, its load profile is barely affected by temporal averaging at 5-minute resolution. However, the typical ON-OFF behavior starts disappearing at 15-minute resolution and is not detectable at 60-minute resolution.

Overall, the load profile of individual electrical appliances appears significantly smoother and loses its characteristics with decreasing temporal resolution. Especially, temporal averaging impacts the maximum power values and the volatility of load profiles. Figure 5.2 quantitatively evaluates this impact on the main types of domestic appliances in the form of box and whisker plots. For that purpose, the maximum power value and the Coefficient of Variation (CV) are computed over the entire measurement period. The CV is the ratio of the standard deviation to the mean and measures the volatility of a time series. In the plot, each data point represents the corresponding indicator value for a certain household, a certain type of appliance, and a certain temporal resolution. The central bar indicates the median value, the box corresponds to the Interquartile Range (IQR), and the ends of the whiskers refer to $1.5 \cdot \mathrm{IQR}$ below and above the lower and upper quartiles, respectively. First of all, a large variance can be seen between different types of appliances but also within the same category. This aspect is not the main focus of interest in this study but is nevertheless consistent with the specifications of such appliances. Moreover, the decrease in temporal granularity systematically leads to a significant drop in both maximal power value and volatility. The load profile of water heaters appears to be the most impacted by the temporal resolution, especially regarding the maximum power value. The impact is particularly pronounced at 60-minute resolution with an average drop of 57% and 38% with respect to 15-minute resolution for the maximum power and the CV, respectively. To a lesser

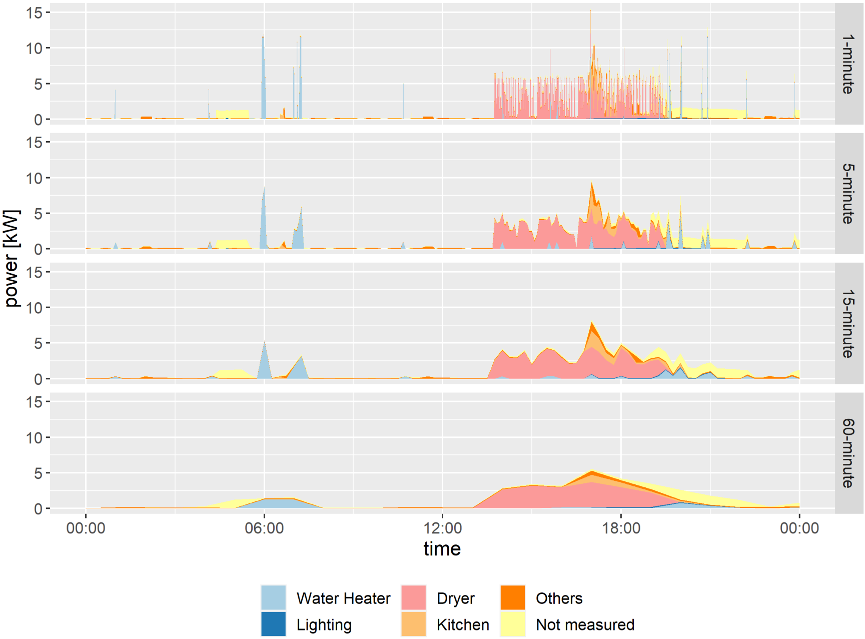


FIGURE 5.3: Load profiling of the residential user presented in 3.9 and displayed at different temporal resolutions.

extent, similar observations can be made for dryers, refrigerators, and kitchen devices. Conversely, lighting is marginally affected, and its load volatility only drops by 13% on average from 1-minute to 60-minute resolution. This is explained by the fact that lighting devices are generally used for longer periods than the measurement time interval.

5.3.2 *Individual Residential End-Users*

The load profile of a household is made of the aggregation of various individual appliances and is likely to be also impacted by the temporal resolution. This is illustrated by Figure 5.3 for a specific residential user. As previously observed, the high and narrow power spikes of the water heater get substantially reduced, and the volatile behavior of the dryer gets smoothed out. At lower temporal resolutions, it is not anymore possible to allocate particular power events to specific electrical appliances. On the one hand, this prevents data-based

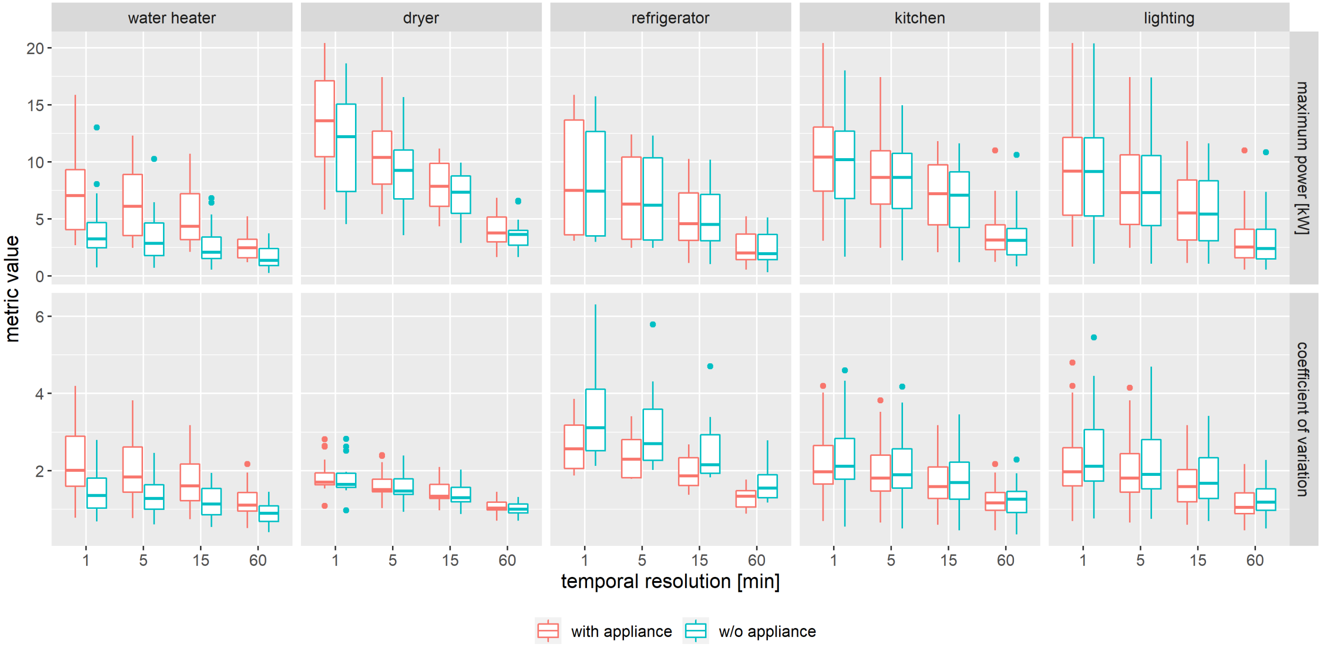


FIGURE 5.4: Box and whisker plots of the maximum value and of the coefficient of variation in active power measurements of residential users at different temporal resolutions. Labels “with appliance” and “w/o appliance” indicate whether the indicators are computed with or without the presence of the specific appliances.

applications (e.g., NILM techniques, estimation of flexibility) from using all necessary pieces of information. In addition, the fact that narrow power spikes appear considerably smaller at lower temporal resolutions is potentially problematic. For example, this can lead to the misestimation of the actual capacity requirements of the local electrical infrastructure. Such concern is highlighted by the authors in [281] in the case of energy storage systems. On the other hand, this ensures a certain level of privacy for the residential user. As noticed in [189], practically only lighting circuits and the refrigerator are still detectable with measurement intervals of 15 minutes. With smart meter data at a 60-minute resolution, only the presence or absence of inhabitants is visible.

Although not explicitly illustrated for the sake of conciseness, temporal averaging still substantially impacts both maximum value and CV at the user level, but to a slightly lesser extent than at the device level. On average, the maximum power value at the user level drops by 17%, 34%, and 66% from 1-minute to 5-minute, 15-minute, and 60-minute resolution, respectively. Analogously, the CV drops by 10%, 24%, and 46% on average from 1-minute to 5-minute, 15-minute, and 60-minute resolution, respectively. Furthermore, Figure 5.4 displays the role of different domestic appliances for both key indicators with respect to the temporal resolution. For that purpose, each data point labeled as “w/o appliance” is based on the load profile of a certain household where the power measurements of the appliance of interest have been subtracted[1]. First of all, the water heater is the appliance that clearly plays the most important role in the household load with respect to both indicators. The dryer also slightly influences the maximum power value, which is not the case of the remaining appliances due to their relatively low power consumption. In contrast to the water heater load, power consumption from the refrigerator as well as the kitchen and lighting devices even lower the volatility of the household load. In addition, there is no substantial impact of temporal resolution on the role of a certain appliance within the total load, which is also valid for the water heater.

5.3.3 *Aggregation of Residential End-Users*

Figure 5.5 finally illustrates the load profiling of the aggregation of all 70 residential users from the Costa Rican sub-metering data set according to different temporal resolutions. This typically corresponds to the loading of

---

[1] Each data point labeled as “with appliance” has a corresponding data point labeled as “w/o appliance” which refers to the same residential user at the same temporal resolution.

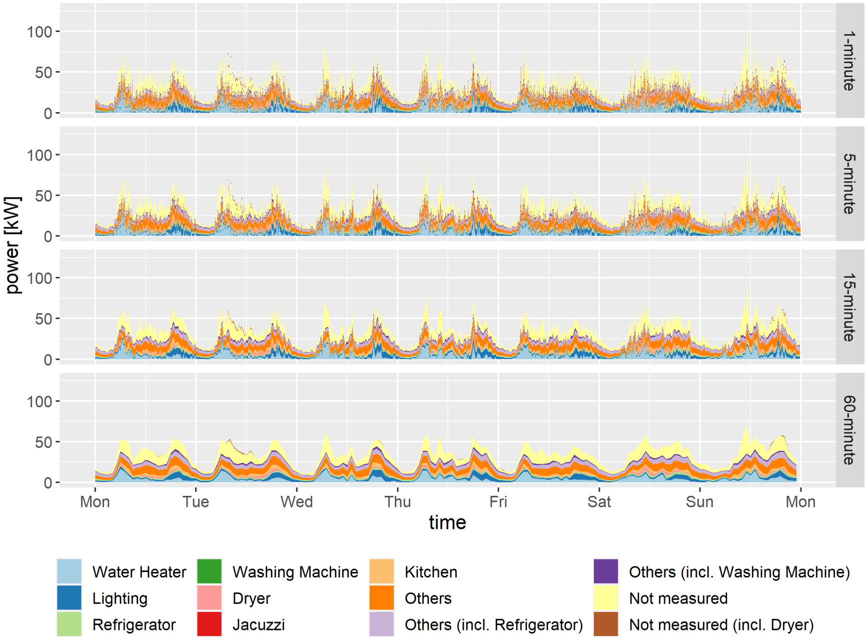


FIGURE 5.5: Load profiling of the aggregation of 70 residential users displayed at different temporal resolutions.

a small distribution transformer. At this spatial aggregation level, only the main tendency is visible. The load of individual appliances or even of specific categories cannot be detected, even at high temporal resolution. Nevertheless, the measurement interval still impacts the maximum power value, which drops by 15%, 27%, and 44% from 1-minute to 5-minute, 15-minute, and 60-minute resolution, respectively. In this case, peak power events are very short (e.g., a few minutes) and get therefore considerably smoothed out by temporal averaging. It must also be noted that such peak power events are not rare and often occur during peak hours. Finally, the volatility is slightly affected as the CV decreases by 3%, 7%, and 14% from 1-minute to 5-minute, 15-minute, and 60-minute resolution, respectively.

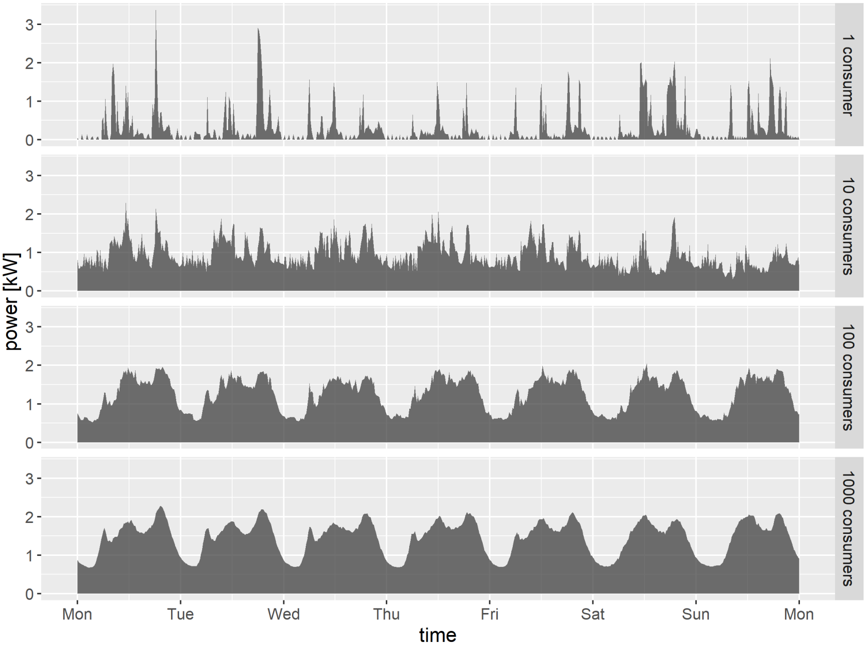


FIGURE 5.6: Smart meter load profiles at different spatial aggregation levels, scaled by the respective number of consumers.

5.4 COMBINED EFFECT OF SPATIAL AND TEMPORAL DIMENSIONS

This section focuses on the effect of spatial aggregation at different temporal resolutions. Spatial aggregation basically refers to the summation of multiple profiles metered at different locations. In this study, the shape and characteristics of individual consumers are compared with the properties of aggregations of 10, 100, and 1'000 consumers. In a distribution grid, the aggregation of 10 consumers is representative of the load visible at an LV bus (e.g., load of a multi-family building). Moreover, the aggregation of 100 and 1'000 end consumers typically corresponds to the loading of an ML/LV transformer and of a small local substation, respectively.

In order to get first insights into the sole effect of spatial aggregation, Figure 5.6 illustrates active power load profiles over one week at different spatial aggregation levels. They come from the Costa Rican smart meter data set with 15-minute measurement intervals. For comparison purposes, profiles are scaled by the number of consumers in each aggregation level. Randomly selected, the upper load profile apparently belongs to a residential consumer. It is characterized by high volatility and no evident periodicity. Each further aggregation level consists of the same consumer(s) as in the previous level, supplemented by new examples from the smart meter data set. The aggregation of 10 consumers partially lowers the volatility of the resulting load profile but does not allow for the emergence of a distinctive pattern. At higher spatial aggregation levels, the load profile gets noticeably less volatile and more periodic. Such characteristics are well-known in the power forecasting community. Indeed, the performance of prediction algorithms significantly improves at higher aggregation levels. This aspect is quantified and confirmed by the authors in [286]. The load profile appears slightly smoother with an aggregation of 1'000 consumers in comparison with 100 consumers, but the main pattern is similar. In this example, it basically tends towards the average typical residential load profile of Costa Rica, as presented in Figure 3.7.

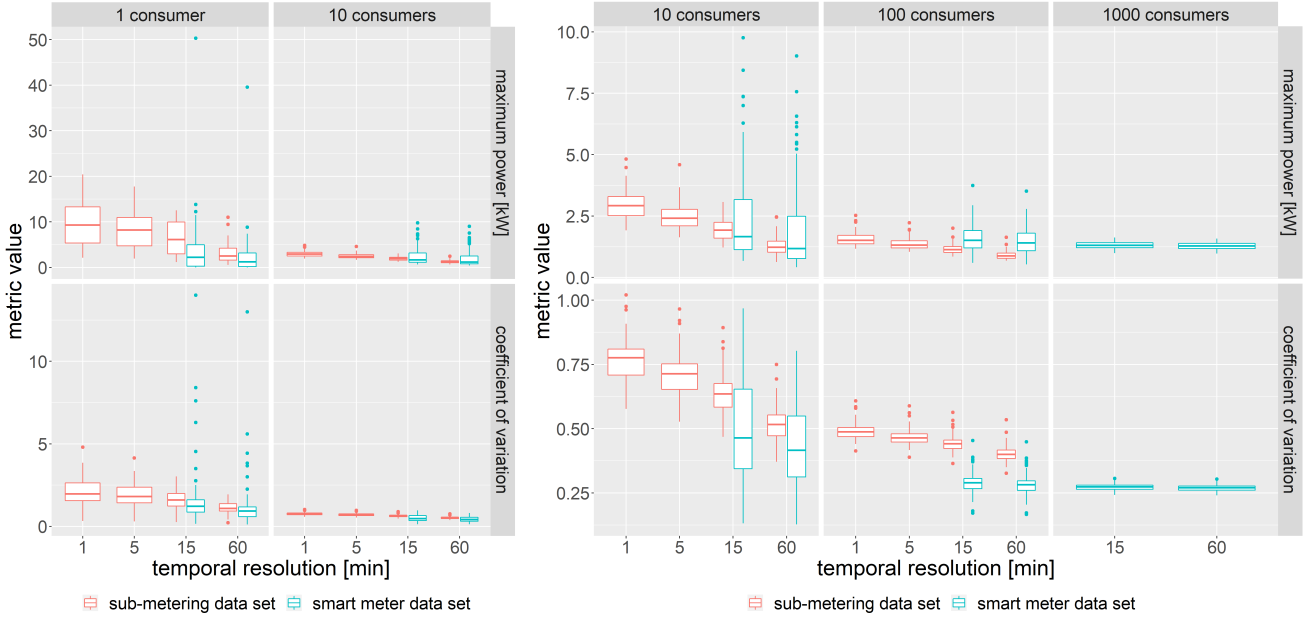
Furthermore, an extensive sensitivity analysis is carried out to quantitatively assess the alteration of active power load profiles with respect to both temporal resolution and spatial aggregation. More concretely, the study considers the combinations of 1-minute, 5-minute, 15-minute, and 60-minute measurement data together with aggregations of 1, 10, 100, and 1'000 end consumers. It must be noted that the number of end consumers is relatively low in the sub-metering data set, while Costa Rican smart meter measurements are not recorded at 1-minute resolution. Therefore, the study is performed on both data sets but is nevertheless restricted to the respective feasible temporal resolutions and aggregation levels. For comparison purposes, aggregate load profiles are again scaled by the respective number of consumers in each aggregation. In order to minimize potential bias in the formation of aggregations, individual load profiles are selected by simple random sampling, which is repeated 100 times. Hence, this leads to the creation of 100 examples of aggregations, each consisting of different load profiles, for each aggregation level. Random sampling is performed without replacement for all cases, with the exception of the sub-metering data set with an aggregation of 100 consumers[2]. For a fair comparison, the same samples per aggregation are maintained over the various temporal resolutions. On this basis, the

---

[2] It must be noted that there are only 70 residential users in the sub-metering data set. Nevertheless, it appears important to also consider aggregations of 100 consumers for this data set. The possibility of sampling more than once the same consumer within a certain aggregation is not seen as problematic at this level, although it must be kept in mind that

Spatial aggregation	Temporal resolution	Scaled maximum power [kW]	Coefficient of variation
1 consumer	1 minute	9.60	2.12
1 consumer	5 minutes	8.01	1.88
1 consumer	15 minutes	6.31 | 3.67	1.58 | 1.58
1 consumer	60 minutes	3.22 | 2.31	1.11 | 1.17
10 consumers	1 minute	2.96	0.77
10 consumers	5 minutes	2.49	0.71
10 consumers	15 minutes	1.95 | 2.48	0.64 | 0.51
10 consumers	60 minutes	1.31 | 2.00	0.52 | 0.43
100 consumers	1 minute	1.56	0.49
100 consumers	5 minutes	1.37	0.47
100 consumers	15 minutes	1.15 | 1.62	0.44 | 0.29
100 consumers	60 minutes	0.90 | 1.49	0.40 | 0.28
1'000 consumers	1 minute	-	-
1'000 consumers	5 minutes	-	-
1'000 consumers	15 minutes	1.31	0.27
1'000 consumers	60 minutes	1.28	0.27

TABLE 5.1: Average value of the indicators of the sensitivity analysis illustrated in Figure 5.7. Red and blue values refer to the sub-metering and smart meter data sets, respectively.



(a) Aggregations of 1 and 10 consumers

(b) Aggregations of 10, 100, and 1'000 consumers

FIGURE 5.7: Sensitivity analysis of the impact of temporal resolution and spatial aggregation on the maximum power value and the coefficient of variation of load profiles.

maximum power value and the coefficient of variation are computed for each load profile and used as indicators.

Figure 5.7 visualizes the outcome of this sensitivity analysis. For the sake of readability, the visualization is split into two sub-figures. Namely, Figure 5.7a displays the characteristics of individual consumers and of aggregations of 10 consumers, and Figure 5.7b focuses on aggregations of 10, 100, and 1'000 consumers. In addition, Table 5.1 summarizes the mean value of both indicators over the 100 repetitions for each feasible combination of temporal resolution and spatial aggregation level. Overall, spatial aggregation leads to the same tendency as temporal averaging, i.e., lower scaled maximum power values and lower coefficients of variation. The largest drop occurs between individual loads and the aggregation of 10 consumers. For example, the scaled maximum power value in the sub-metering data set is approximately divided by three at all temporal resolutions. The CV is also divided by three in both data sets when aggregating ten consumers, almost independently of the temporal resolution. For both indicators, not only the average value but also the variance[3] among the 100 different repetitions at the same level noticeably decreases with increasing aggregation level. Moreover, previous analysis on the sole impact of the temporal resolution appears to be generally valid at any spatial aggregation level. Notably, the transition from 15-minute to 60-minute resolution causes the most significant decrease of both indicators, although the impact gradually fades with increasing spatial aggregation level. Finally, the influence of temporal averaging is practically null for aggregations of 1'000 consumers.

By comparing the sub-metering and smart meter data sets, the larger variety of consumers in the latter is confirmed, especially in light of the various outliers at low aggregation levels. With the exception of those outliers, the scaled maximum power values in the smart meter data set appear, on average, less sensitive to the aggregation level than in the sub-metering data set. Such differences are explained by the fact that the smart meter data set does not only consist of volatile residential load profiles but also of other loads with inherently lower variability.

---

the diversity is limited. Conversely, aggregations of 1'000 consumers based on 70 examples would be totally biased and are therefore not considered in the sensitivity analysis.

[3] The variance is reflected in the width of the IQR.

5.5 CONCLUSION

To conclude, this chapter clearly demonstrates that the shape and statistical properties of AMI load profiles are substantially impacted by the choice of the measurement interval and by the aggregation level at which they are metered. This generally holds true for aggregations smaller than 1’000 consumers, which is typically the case in distribution grids. First, temporal resolution plays a significant role when metering individual domestic appliances. Per definition, temporal averaging does not affect the average energy consumption but considerably influences the load shape as well as key features like the maximum recorded power and the coefficient of variation. Hence, a decrease in temporal resolution is generally associated with considerable loss of information for individual appliances. The effect is even more pronounced for high power consumption appliances (e.g., water heater, dryer) whose impact on the overall system is non-negligible. The same observations are visible at the level of a residential user and of an aggregation of users, but to a slightly lesser extent. For example, the aggregation of 70 users as presented in Section 5.3.3 corresponds to the loading of a small ML/LV transformer. Even at this aggregation level, power measurements at lower temporal resolutions fail to account for short peak power events that might be of importance for transformer sizing. In comparison with smart meter data, measurements at the transformer level are not particularly critical in terms of privacy or size of data. Hence, high-frequency data are of great interest to faithfully represent the actual loading situation of transformers.

Furthermore, observations at individual aggregation levels are confirmed by the sensitivity analysis. Based on two measurement data sets with different characteristics, the analysis covers a large range of temporal resolutions and spatial aggregation levels which are common in distribution grids. The same general trends are observed in both data sets, where load profiles are largely smoothed out at lower temporal resolutions and higher spatial aggregation levels. The impact is particularly significant when aggregating ten consumers. To a lesser extent, the maximum power value and the coefficient of variation are gradually reduced with increasing measurement time intervals.

Despite the important role of temporal and spatial dimensions on AMI measurements, they are rarely explicitly considered in data-based studies and in the development of data-based algorithms at the level of distribution grids. It is however known that peak power events can be harmful to power systems. Their proper estimation is crucial for the infrastructure’s dimensioning. Moreover, lower volatility especially allows for better predictability, which is a key aspect for many data-based algorithms, e.g., forecasting models. The modification of the inherent properties of measurement data because of temporal averaging or spatial aggregation might lead to biased conclusions which are not in phase with reality. Hence, the following chapters shall take these concerns into account in AMI data-based applications. Notably, distribution grid operators use to approximate the load profile of single consumers without a smart meter based on the load measured at a higher spatial aggregation level (e.g., at the feeder level). Chapter 7 thoroughly studies the implications of such approximation for the estimation of voltages and power flows. Furthermore, the impact of temporal granularity on the success of load disaggregation algorithms is detailed in Chapter 8.

Part II

DATA-BASED MODELING OF DISTRIBUTION GRIDS

6

PSEUDO-MEASUREMENT SYNTHESIS

This chapter focuses on creating active and reactive power pseudo-measurement for end-consumers in low-voltage grids. At this level, the coverage of advanced meters is insufficient to obtain an observable system, which necessitates the synthesis of pseudo-measurements. In the first stage, novel load profile generation approaches for both active and reactive power are proposed based on existing smart meter data and compared to traditional techniques. The proposed approach for active power synthesis is stochastic and ensures realistic statistical properties at the individual level, which is generally ignored in the context of pseudo-measurement synthesis. Rarely considered in the literature, the synthesis of reactive power is also thoroughly addressed. In the second stage, a methodology is developed to optimally allocate synthetic load profiles to actual non-metered consumers. On the one hand, the spatial aggregation of all load profiles must coincide with power measurements at an aggregate level. On the other hand, the energy contained in each allocated load profile must match the reported energy consumption of the corresponding end-user. All presented approaches are evaluated in different case studies using smart meter data from Costa Rica and from the City of Basel. This chapter is essentially based on [288].

6.1 INTRODUCTION

In traditional power systems, LV grids were regarded as a black box in terms of information. The power was known to flow from higher grid levels to the end-consumers, and the grid infrastructure was usually over-dimensioned to cope with the most probable contingencies. For proper operation of distribution grids, it was sufficient to model only the MV level and aggregate all loads at the MV/LV feeders. At this level, the use of standard load profiles allowed for satisfactory representation of the system in the absence of actual measurements. However, current DSOs face new operation and control challenges due to the increasing share of renewable energy sources, distributed storage units, and electric vehicles. In this new paradigm, power flows are not unidirectional anymore, and the emergence of new load types provokes previously unimaginable grid congestions and voltage instabilities in LV grids. Efficiency is also a key aspect of a sustainable system, which requires the development of alternative solutions to traditional grid reinforcement to cope with contingencies. Hence, properly modeling and monitoring distribution grids down to the lowest voltage level, instead of only relying on aggregated data, has become more important in recent years. This step highly relies on the information provided by smart meters at the level of end-consumers.

Nevertheless, a large majority of the DSOs have not reached a full smart meter penetration in their system yet, as discussed in Section 2.3. For many countries, especially in Europe, North America, and the Asia-Pacific area, the partial availability of smart meter data is a transitory state, and a wide-scale coverage should be achieved within the next decade. In contrast, a few states have opted for a selective smart meter roll-out due to cost considerations and stronger data privacy policies. Moreover, the installation of smart meters does not necessarily imply the full availability of their data. Smart meters and the corresponding communication infrastructure are not infallible, and a small portion of the potential data in the system are often not exploitable. Besides, the recording of high-resolution power measurements at the end-user level is associated with serious privacy concerns. This considerably restricts their usage outside the limited scope of applications defined by privacy protection laws. In summary, although an increasing share of end-users are equipped with smart meters, their data are still a scarce commodity in the power system community. In any case, the amount of reliable data in LV grids is insufficient to obtain an observable system.

The synthesis of pseudo-measurements appears as an indispensable step towards increasing the transparency in LV grids. They allow for more robust decision-making processes in terms of grid operation and control, but also long-term planning. Notably, pseudo-measurements are widely used in Distribution System State Estimation (DSSE) [114]. In fact, a robust DSSE requires an observable system and a certain redundancy of measurements which can only be achieved with the help of pseudo-measurements. Extensive literature exists for that purpose, which principally focuses on active power. Pseudo-measurements synthesis techniques have first been developed for the transmission network before being adapted to distribution grids and finally down to the load of individual end-users. The conception of synthetic load profiles for end-consumers is generally inspired by the models used at higher spatial aggregation levels, where the load is particularly smooth and periodic. However, the load profile of individual consumers is highly volatile, and its shape and characteristics can considerably vary among different consumers, as seen in Chapter 5. Unfortunately, such features are often not reflected by state-of-the-art pseudo-measurement synthesis techniques that are simply designed to minimize the state estimation error at the system level. Alternatively, different load profile generators are suggested in the literature outside the DSSE context. They usually generate very realistic load profiles at the level of individual residential consumers. Nevertheless, there are computationally intensive and require detailed information to faithfully represent a certain type of end-user, which is not applicable at the level of a distribution grid.

This chapter focuses on the synthesis of realistic pseudo-measurements at the level of end-users while complying with aggregate information in a given LV grid. The idea is to leverage only smart meter data and aggregate pieces of information which are typically available in distribution grids. Section 6.2 details the methodology behind the synthesis of pseudo-measurements. Existing synthesis approaches are reviewed, and a novel approach based on Markov Chain Models (MCMs) is proposed for the synthesis of active power profiles. Although the use of MCMs is not new in this context, they traditionally require a substantial amount of training data for the creation of statistically representative load profiles. The innovation lies in the design of the transition matrix, which inherently accounts for variations over time without excessive training data. Besides, the section also discusses the synthesis of reactive power, which is rarely addressed in the literature. Concretely, the relationship under various conditions between active and reactive power consumption is analyzed, which leads to the development of a novel reactive power synthesis approach. Furthermore, synthetic load profiles are optimally allocated to non-metered consumers in a distribution grid. For that purpose, different optimization problems are solved in order to best match synthetic load profiles with power measurements at a spatial aggregate level, but also with the energy requirements of each individual consumer. Section 6.3 evaluates the effectiveness of the different approaches presented in Section 6.2. Active and reactive power synthesis are tested on the smart meter data from the City of Basel and from Costa Rica, respectively. Finally, the main contributions of this chapter are summarized, and future work is outlined in Section 6.4.

6.2 METHODOLOGY

This section presents the proposed methodology to synthesize active and reactive load profiles for non-metered consumers in a distribution grid. Syn-

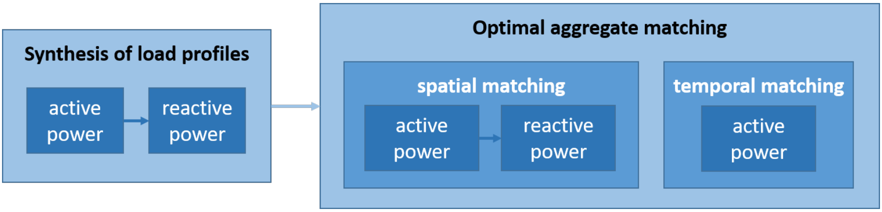


FIGURE 6.1: Procedure for the synthesis of active and reactive power pseudo-measurements.

thetic load profiles must make sense both at the level of individual consumers and at an aggregate level. In terms of data availability, the methodology relies on three main assumptions that are reasonable in current distribution grids. First, the grid is characterized by partial smart meter penetration, meaning that active and possibly reactive power measurements are recorded for a portion of end-users. As argued above, partial smart meter penetration can either be a transitory or permanent state. Second, the average energy consumption of non-metered consumers is known. This is anyway required for billing purposes, where traditional meters are typically read every month or every year, depending on the local practices. Third, power measurements are recorded at an aggregate level. This holds true in local substations for monitoring purposes and tends to be more and more common at the level of distribution MV/LV transformers.

Figure 6.1 summarizes the procedure followed in this chapter to create active and reactive power pseudo-measurements. First of all, active power profiles are synthesized as presented in Section 6.2.2, leveraging existing smart meter data or aggregate measurements. The synthesis of reactive power profiles is directly based on active power via the power factor, which is addressed in Section 6.2.3. Both sections start with a literature review before describing several synthesis techniques. Among others, a standard approach traditionally used by DSOs and a novel and more realistic approach are detailed for both active and reactive power synthesis. Subsequently, Section 6.2.4 defines an innovative procedure to optimally allocate synthetic load profiles to actual non-metered consumers while being consistent at an aggregate level. Based on partial smart meter penetration, a power gap results from the difference between the power profile measured at an aggregate level (e.g., measurements of transformer loading) and the spatial aggregation of all metered end-users' profiles. The optimal aggregate matching procedure starts with the selection of the most appropriate examples from a large set of synthetic profiles to fill the active power gap. Next, corresponding synthetic reactive load profiles are scaled to perfectly match the reactive power gap. Finally, selected load profiles are assigned to individual end-users according to their reported active energy consumption.

6.2.1 *Power Gap Profile*

In the case of partial smart meter penetration, the load profile of individual non-metered consumers is not known. Nevertheless, the fusion of smart meter data with measurement data at an aggregate level (e.g., distribution transformer, local substation) provides aggregate information about the fraction of the load which is not metered in a distribution grid. More precisely, an active power gap profile results from the difference between the active power profile measured at the aggregate level and the spatial aggregation of all metered customers' profiles:

$$P_{\mathrm{gap},t} = (1 - \rho_P) \cdot P_{\mathrm{agg},t} - \sum_{j=1}^{n} P_{j,t}, \quad \forall t \in \{1, 2, \ldots, T\}, \tag{6.1}$$

where $P_{\mathrm{gap},t}$, $P_{\mathrm{agg},t}$ and $P_{j,t}$ are the active power values at time t of the gap profile, at the aggregate level and of metered consumer j, respectively. The parameter ρ_P models the active power losses in the distribution grid, n is the number of consumers equipped with a smart meter, and T is the total number of time steps under consideration. Such active power gap is illustrated in Figure 3.7 in the case of a sub-grid in the City of San José. In this example, the portion of the total load labeled as "non metered" primarily belongs to non-metered consumers, but also partly comes from losses in power lines and transformers. These losses are first given a rough average estimate (e.g., between 1% and 5% of the total load depending on the distribution grid infrastructure) and can be determined precisely in a later stage by load flow simulation or state estimation, as presented in Chapter 7.

Analogously, a gap profile also appears for reactive power:

$$Q_{\mathrm{gap},t} = (1 - \rho_Q) \cdot Q_{\mathrm{agg},t} - \sum_{j=1}^{n} Q_{j,t}, \quad \forall t \in \{1, 2, \ldots, T\}, \tag{6.2}$$

where $Q_{\mathrm{gap},t}$, $Q_{\mathrm{agg},t}$ and $Q_{j,t}$ are the reactive power values at time t of the gap profile, at the aggregate level and of metered consumer j, respectively. The parameter ρ_Q models the reactive power losses in the distribution grid.

6.2.2 *Synthesis of Active Power Profiles*

The synthesis of active power profiles is a relatively well-established topic in the power system literature. Nevertheless, the increased attention given to distribution grids considerably boosted the need for pseudo-measurements due to the obvious lack of actual measurements. Traditionally, two main categories are distinguished for the synthesis of active power load profiles: bottom-up and top-down approaches.

Bottom-up approaches rely on the modeling of single appliances. In this case, the synthetic load profile of end-users consists of the aggregation of all appliances' load. The resulting load profile usually depends on the dwelling characteristics, on the type and specifications of appliances, and possibly on behavioral models of consumer habits. For example, LoadProfileGenerator (LPG) is a well-recognized tool to generate residential load profiles based on a set of predefined appliance models and on a full behavior simulation of inhabitants [218]. Similarly, the approaches developed in [289] and [290] transforms user activity profiles into load models based on a Markov Chain Monte Carlo (MCMC) method and on appliance ownership statistics. Alternatively, the authors in [291] accurately model the thermal, electrical behavior of domestic appliances to capture the time-variant response of residential consumers to changes in voltage. Bottom-up approaches create very realistic load profiles but also have major downsides. First, existing tools only focus on residential loads that are generally made of well-defined appliances, in contrast to commercial and industrial loads. In addition, detailed information on the appliances (e.g., type and number of appliances, temperature set point), the end-users (e.g., household composition, number of inhabitants), and the buildings (e.g., type of house, number of stories, insulation material) is often required to faithfully model the loads in a specific system. Nevertheless, such statistical data are often not available, notably due to privacy reasons. Furthermore, modeling the characteristics of every single appliance and the behavior of every single end-user is computationally intensive. This limits the generation of a good diversity of load profiles and prevents the application to large populations.

Top-down approaches make use of actual measurement data to generate new profiles with similar properties. They are often leveraged in the DSSE literature, where actual measurements are not sufficient to obtain an observable system. Standard Load Profiles (SLPs) are widely used for that purpose. An SLP corresponds to the average power consumption profile of a specific class of end-consumer. The outcome is similar to the average profiles resulting from the cluster analysis presented in Section 4.4.1. For example, the authors in [292] make use of SLPs of four typical classes in the UK to generate active power pseudo-measurements. Since pseudo-measurements inspired by SLPs reflect the general trend of an ensemble of consumers, their shape is particularly smooth and periodic. Hence, they are appropriate for approximating the load at an aggregate level. However, the authors in [293] point out that pseudo-measurements based on Standard Load Profiles (SLPs) are not satisfying for DSSE in LV grids where the load is highly volatile and hardly predictable. This averaging or smoothing effect is also induced by other approaches used for the generation of pseudo-measurements. Some examples suggested in the literature rely on Gaussian Mixture Models (GMMs), Artificial Neural Networks (ANNs), clustering techniques, autoregressive models, or Gradient Boosting Tree (GBT) models [292, 294–297]. Especially, supervised algorithms are trained to minimize the point-wise error with the original data and fail to reflect the original load distribution, similarly to SLPs. More detailed information on this smoothing effect is given in Section 9.3. Alternatively, Generative Adversarial Networks (GANs) are recently gaining in popularity due to their great capacity to generate synthetic data with similar statistical properties as real data. Notably, the authors in [298] show that the load distribution of residential pseudo-measurements created by GANs is faithful to reality. Nevertheless, GANs require a substantial amount of training data and are difficult to interpret due to their black-box nature, which is the case of most ANN-based models.

In this work, the synthesis of active power pseudo-measurements is based on Markov Chain Models (MCMs) which are leveraged as top-down approaches. These discrete-time stochastic processes are particularly effective at faithfully representing the sharp spikes and high volatility of load profiles at the end-user level. Besides, they do not necessitate an extensive amount of input data. In the following, a method called “Standard Load Allocation” and traditionally used by power utilities is first presented as a benchmark. Subsequently, a traditional MCM-based approach is detailed according to state-of-the-art literature. On this basis, the author suggests a novel version called “Adaptive Markov Chain Model” which inherently accounts for seasonality at different time scales without excessive computational effort. This adaptive version has first been presented in [288].

6.2.2.1 *Standard Load Allocation*

The Standard Load Allocation (SLA) is a pseudo-measurement synthesis approach that only leverages power measurements at an aggregate level and the average energy consumption of individual non-metered consumers. Especially, it does not involve smart meter data. Due to its simplicity, the approach is often used by DSOs in the absence of smart meter data and is therefore chosen as a benchmark in this work. Developed by the authors in [299], the SLA consists of assigning the power gap profile as defined in Equation (6.1) to all non-metered end-users, scaled by their respective share of energy consumption:

$$P_{i,t} = \frac{E_i}{\sum_{j=1}^{n_0} E_j} \cdot P_{\mathrm{gap},t}, \quad \forall i \in \{1, 2, \ldots, n_0\},\ \forall t \in \{1, 2, \ldots, T\}, \tag{6.3}$$

where $P_{i,t}$ is the active power allocated to consumer i at time t, $P_{\mathrm{gap},t}$ is the active power value of the gap profile at time t, and E_i is the average energy consumption of consumer i. In addition, n_0 is the number of non-metered consumers, and T is the total number of time steps under consideration.

Since the SLA is based on measurements at a higher spatial aggregate level, synthetic load profiles also inherit from their statistical properties, e.g., reduced volatility compared to actual smart meter data. Statistically speaking, the load distribution of SLA-based pseudo-measurements is similar to the load distribution of SLPs or, to a certain extent, load profiles generated by algorithms focusing on the point-wise error (e.g., autoregressive models, clustering algorithms, most supervised ML approaches).

6.2.2.2 *Traditional Markov Chain Model*

A Markov model is particularly suitable for modeling systems where the current state of a sequence is highly correlated to the immediately preceding state, which is the case of load profiles. The model itself consists of a finite set of states and transition probabilities among these states. It builds on existing data and allows for the generation of time series with similar properties. In the context of power measurements, a state corresponds to a range of power consumption values as observed in the input data. In order to define the states, different approaches are proposed in the literature. In [300], the entire range of possible power values is divided into segments of equal length. However, this leads to the formation of states with very few observations, usually corresponding to higher power values for smart meter data. This could bias the underlying distribution of high power values due to the lack of corresponding data. Alternatively, the authors in [301] define the limits of the states by equally splitting the cumulative probability density function. Hence, all states contain the same number of observations but refer to power ranges of different lengths. Notably, this leads to less granular modeling of higher power values. In this work, the definition of the states is based on K-Means clustering, where the power values of the input time series are used as training data. As a result, each cluster defines a state which corresponds to a different range of power. This allows for a trade-off between the creation of power ranges of the same length and states with the same number of observations. A compromise must also be found regarding the number of states. Although a high number is desired for a more detailed model, there is the risk of creating an overfitted model. Indeed, more states imply fewer input data to compute the probabilities for each transition among the states.

Once the range of each state is defined, the probability $p_{r \to s}$ to jump from a certain state r at time step t to another state s at time step $t+1$ is statistically derived from the input time series. This leads to the creation of a so-called transition probability matrix of dimension $n_s \times n_s$, where n_s is the total number of states. Each element of this matrix reflects the relative probability of moving from the current state to any other state, including the current state. Rows and columns of the transition matrix refer to the current and next state, respectively. In a first-order Markov Chain Model (MCM), the transition probability to a certain state at time t only depends on the state at time $t-1$. In contrast, a q^{th}-order MCM considers the states of q previous time steps. In this work, only the first-order MCM is implemented. A higher-order MCM would substantially reduce the amount of data available for the calculation of each transition probability.

An MCM is said to be time-homogeneous if the transition matrix remains constant over time. However, it has been observed that the load distribution of electricity end-users can significantly vary over time, especially at different periods of the day. The authors in [302] show that a time-homogeneous MCM fails to model the effect of the time of the day on electricity consumption patterns. Hence, a time-inhomogeneous model is implemented, where different states and various probability factors are defined for each time step within a day. Assuming a 15-minute temporal resolution, this leads to a transition matrix of dimension $96 \times n_s \times n_s$. In this case, distinct subsets from the original set of power observations are used to calculate each probability factor. For example, only power values measured at 10:00 within the load range of the first state are used to calculate the probabilities of moving from the first state at 10:00 to any other state at 10:15. Per definition, all relative probabilities must sum to one for each row of the transition matrix (i.e., each starting state).

Finally, the generation of a synthetic load profile using a Markov chain occurs time step after time step by a random walk through the transition probability matrix. In other words, starting from a predefined state r, the next state s is repeatedly selected according to the probability $p_{r \rightarrow s}$ given by the element at row r and column s in the transition matrix. By nature, a Markov chain produces a sequence of discrete states. To obtain continuous values in this version of the traditional MCM, each state (i.e., load range) is further divided into a fixed number of sub-levels associated with a relative probability that approximates the load distribution within the load range. This can be interpreted as the creation of a density histogram for each state, where power values are selected. Eventually, white noise is added to the power values to make the synthetic profile more realistic. In theory, one model allows for the creation of an infinite number of synthetic profiles with similar transition probabilities as for the input load profile.

6.2.2.3 *Adaptive Markov Chain Model*

As it will be seen in Section 6.3.1, the traditional MCM cannot represent seasonality on a medium- and long-term basis unless it is explicitly designed for this purpose. However, such an explicit design often implies a loss of model accuracy due to overfitting. Indeed, the more temporal distinctions are drawn in a time-inhomogeneous model, the fewer observations are available to calculate the probabilities for each specific period of time. In order to overcome this situation, a novel approach called "Adaptive Markov Chain Model" (AMCM) is proposed. It generalizes the concept of time-inhomogeneity without loss of accuracy or need for additional input data. Concretely, each element of the original $n_s \times n_s$ transition matrix is not a real number between 0 and 1 anymore, but a logistic regression model that learns the corresponding transition probability:

$$h_\theta(x) = g\left(\theta^{\mathrm{T}} x\right), \quad \text{with } g(z) = \frac{1}{1+e^{-z}}, \tag{6.4}$$

where $h_\theta(x)$ is a logistic regression model[1], $x = (1, \mathit{hour}, \mathit{weekday}, \mathit{month})$ is a vector of time-related features, and $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)$ a vector of coefficients defined by the training process. In practice, the output of the logistic regression is often set to 0 or 1 if it is lower or higher than 0.5, respectively. This final step is however not taken in this case. Hence, depending on the hour of the day, the weekday, and the month, the function outputs different transition probabilities lying between zero and one. For a certain starting

---

[1] To be precise, the element describing the transition probability from state r to s should be defined as $h_{\theta, r \rightarrow s}$. Subscript $r \rightarrow s$ is nevertheless discarded for the sake of simplicity.

state r, the training process of the corresponding logistic regression model aims to minimize the cost of prediction error:

$$J_r(\theta) = -\frac{1}{|\mathcal{T}_r|} \cdot \sum_{t \in \mathcal{T}_r} \Big[y_i \cdot log\left(h_\theta(x_i)\right) + (1 - y_i) \cdot log\left(1 - h_\theta(x_i)\right) \Big], \quad (6.5)$$

where $J_r(\theta)$ is the cost function of the logistic regression model with the set of coefficients θ referring to the starting state r. In addition, $h_\theta(x_i) \in [0, 1]$ is the transition probability function (i.e., logistic regression model) evaluated at time i, and $y_i \in \{0, 1\}$ indicates whether the corresponding transition actually occurs at time i. Eventually, $\mathcal{T}_r$ is the set of all time steps related to state r.

So far, the training for each element of the adaptive transition matrix has been presented as a binary logistic regression with only two possible outcomes (i.e., either the transition occurs or not). Nevertheless, the rows and columns of the adaptive transition matrix are linked and refer to a certain starting and landing state in the transition process, respectively. In fact, each row can be implemented as a multinomial logistic regression with multiple possible outcomes (i.e., either the transition occurs to state 1, or to state 2, or to state 3, etc.). Per definition, the multinomial logistic regression model is designed such that the whole set of probabilities forms a probability distribution. This means that all probabilities in a certain row of the transition matrix theoretically sum up to one for any predefined set of features. This is in line with the definition of the transition matrix of a Markov chain.

In this way, the AMCM is able to learn from the input data and reproduce the notion of seasonality on several time scales. Knowing that $24 \times 7 \times 12 = 2'016$ combinations of hours, weekdays and months exist, the AMCM can also be seen as a traditional MCM whose transition matrix has a dimension of $2'016 \times n_s \times n_s$. In contrast to the traditional MCM, resulting probabilities are nevertheless estimated based on the whole set of available data, which mitigates the risk of overfitting. It should also be noted that this model can be easily fine-tuned by making use of different time-related features and by including exogenous variables (e.g., temperature).

Once a sequence of discrete states is obtained from the AMCM, the conversion into actual power values is accomplished by means of a continuous probability function. Instead of discrete sub-levels as presented for the traditional MCM, the probability function directly models the continuous load distribution. Different ways of approximating a load distribution are mentioned in the literature, e.g., Weibull, Log-normal, Beta, Gamma, Gaussian, and Generalized Extreme Value distributions [300, 303, 304]. However, the authors in [305] show that power measurements in distribution systems usually do not follow a specific probability density function. They suggest the Gaussian Mixture Model (GMM) as the most appropriate approach to represent different types of unknown load distributions. Formally, a GMM consists of the linear combination of several Gaussian distribution components:

$$GMM(y) = \sum_{g=1}^{K} \omega_g \cdot f_g(y), \quad \text{with } f_g \sim \mathcal{N}(\mu_g, \sigma_g^2), \tag{6.6}$$

where $GMM(y)$ is the GMM evaluated at value y, $f_g(y)$ is the g^{th} Gaussian component evaluated at value y, and ω_g is a weighting factor associated with $f_g(y)$. In addition, K is the number of Gaussian componants, and $\mathcal{N}(\mu_g, \sigma_g^2)$ refers to the Gaussian (or normal) distribution defined by its mean μ_g and variance σ_g^2. More precisely, each Gaussian components is defined as follows:

$$f_g(y) = \frac{1}{\sqrt{2\pi \cdot \sigma_g^2}} e^{\frac{-(y-\mu_g)^2}{2\sigma_g^2}}. \tag{6.7}$$

The Expectation-Maximization (EM) algorithm is one of the best data-based approaches to finding the maximum-likelihood estimate of the parameters (i.e., weight, mean, and variance) for each of the K Gaussian components [306]. This is an iterative method that is first initialized with K standard Gaussian components randomly placed in the data space. Next, the following two steps are repeated until convergence:

1. *Expectation*: According to its location in the data space, each data point is given a probability to belong to each of the Gaussian components.
2. *Maximization*: The parameters of each Gaussian component are re-estimated to best fit all softly assigned data points (i.e., data points weighted by their given probabilities).

The ideal number of Gaussian components can be derived from the Akaike Information Criterion (AIC). This criterion finds a trade-off between the complexity of a statistical model and the maximum value of its likelihood function:

$$AIC = 2 \cdot p - 2 \cdot ln(\hat{L}), \tag{6.8}$$

where p is the number of model parameters to estimate and $\hat{L}$ is the maximum value of the model's likelihood function. In practice, the number of Gaussian components is increased until the difference between K and $K+1$ components is below 0.5%. More information is given in [306].

In order to generate continuous values from the discrete states in the Markov chain, the authors in [300] propose to build a GMM on each subset of input data related to the respective states. However, the load range of a single state is often too narrow to properly fit a GMM, which leads to a substantial mismatch between the input data and the load distribution model. In this work, one GMM per hour of the day is calculated over the entire range of power consumption. This is justified by the fact that the load distribution highly varies over different periods of the day. Therefore, the power value at a certain time step of the synthetic load profile is generated by randomly sampling the hour-related GMM. Note the sampling still occurs only within the load range defined by the state.

6.2.3 *Synthesis of Reactive Power Profiles*

The reactive power of a load is basically linked to its active power consumption via the power factor:

$$Q_t = P_t \cdot \tan(\arccos(pf_t)), \tag{6.9}$$

where Q_t, P_t, and pf_t correspond to the reactive power, active power, and power factor at time t. A power factor value of one implies the absence of reactive power. For a fixed active power value, note that the relationship between the power factor and the reactive power is highly non-linear.

Reactive power pseudo-measurements are notably required in load flow simulation and DSSE, where an observable system is required. Among others, reactive power influences key quantities of distribution grids, such as the voltage or the loading of grid components (e.g., distribution lines, transformers). Hence, their synthesis must be properly handled in the absence of actual measurements. Unfortunately, the literature on pseudo-measurement synthesis largely neglects reactive power. When it is considered, reactive power is commonly created based on active power by assuming a fixed average power factor [292, 297]. In rare cases, reactive power is synthesized based on the same method as for active power [293, 294]. This leads to the same considerations as previously discussed for active power synthesis, especially regarding the smoothing effect.

In this work, the synthesis of reactive power profiles is based on the estimation of the power factor. Hence, active power measurements or pseudo-measurements must be already available. Often assumed by DSOs and in literature, the fixed average power factor is used as a benchmark model. As an alternative, the power factor of all consumers is modeled as being equal to the time-variant power factor observed at an aggregate level. Finally, an innovative approach is proposed, which considers the power factor as a time-variant quantity that depends on different observable features at the level of individual end-consumers. This novel approach has been partially developed in [307] and is briefly presented in [308].

6.2.3.1 *Average Power Factor*

In the absence of high-resolution measurements, the power factor is traditionally modeled as a time-invariant quantity. For example, its value is often set to the average power factor value observed in the system at a higher spatial aggregation level. Depending on the type of distribution grid, DSOs generally estimate the average power factor between 0.95 and 0.99 inductive. Consequently, each consumer without reactive power measurements is assigned with the same average power factor:

$$pf_{i,t} = pf_{\text{avg}}, \quad \forall i \in \{1, 2, \ldots, n_0\},\ \forall t \in \{1, 2, \ldots, T\}, \tag{6.10}$$

where $pf_{i,t}$ is the power factor allocated to consumer i at time t, and pf_{avg} is an average power factor estimated by the system operator. In addition, n_0 is the number of consumers without reactive power measurements, and T is the total number of time steps under consideration. On this basis, synthetic reactive power profiles exhibit the exact same shape as the respective active power profiles, scaled by $\tan(\arccos(pf_{\text{avg}}))$.

6.2.3.2 *Aggregate Power Factor*

Similar to the standard load allocation for active power synthesis, this second approach relies on information at the aggregate level. More precisely, the so-called aggregate power factor can be derived from the active and reactive power quantities at the aggregate level and applied to all consumers without reactive power measurements at all time steps:

$$pf_{i,t} = pf_{\text{agg},t}, \quad \forall i \in \{1, 2, \ldots, n_0\},\ \forall t \in \{1, 2, \ldots, T\}, \tag{6.11}$$

where $pf_{\text{agg},t}$ is the power factor measured at the aggregate level at time t. Remaining variables and parameters are defined as in Equation (6.10). Hence, the power factor is modeled as time-variant quantity, but all consumers exhibit the same relationship between active and reactive power.

6.2.3.3 *Adaptive Power Factor*

The proposed approach accounts for the fact that the power factor is a time-variant quantity that certainly differs across various consumers. To be applicable, it assumes partial smart meter penetration, where both active and reactive power quantities are recorded. In the case where smart meters only provide active power measurements, a data set (e.g., open data set) with active and reactive power measurements from similar consumer types should at least be available. The idea is to gain insight into the variation of the power factor based on available smart meter data. Subsequently, adaptive power factors are generated and applied to the end-users where only active power (pseudo-)measurements are available in order to synthesize reactive power pseudo-measurements.

To begin with, Figure 6.2 illustrates reactive with respect to active power values for four different consumers from the Costa Rican smart meter data set[2]. Power measurements are distinguished by color according to the period of the day when they have been recorded, i.e., morning (5:00-9:00), day (9:00-17:00), evening (17:00-21:00), and night (21:00-5:00). It is undeniable that the relationship between active and reactive power substantially varies across different consumers. Nevertheless, some trends are visible. First, the reactive power of small consumers (i.e., end-users with a relatively low average active power consumption) appears at two distinct levels for the same active power value, which is dispersed for large consumers. Besides, larger consumers are likely to consume more reactive power for the same level of active power. Second, power data points tend to be located in different regions of the scatter plot depending on the measurement period within the day. Third, the general relationship of reactive power to active power is not always linear and tends to flatten out at higher active power levels, especially for small consumers.

Accordingly, the power factor seems to vary in different conditions. In fact, the power factor mainly depends on the type of appliances in use by each consumer at each time step. As illustration, Table 6.1 summarizes the power factor values of domestic appliances reported in [287]. The power factor basically varies from 0.44 for a clothes dryer to 1 when pure heating is involved. It does not only vary across different electrical appliances but also for various modes or cycles of a specific appliance. In any case, most of the appliances cannot be detected based on standard smart meter data, as noticed by the authors in [189]. Nevertheless, different observable features

---

[2] In this example, small consumers have an average power consumption of 0.42 kW and 0.93 kW, which rises to 3.82 kW and 4.17 kW for the large consumers.

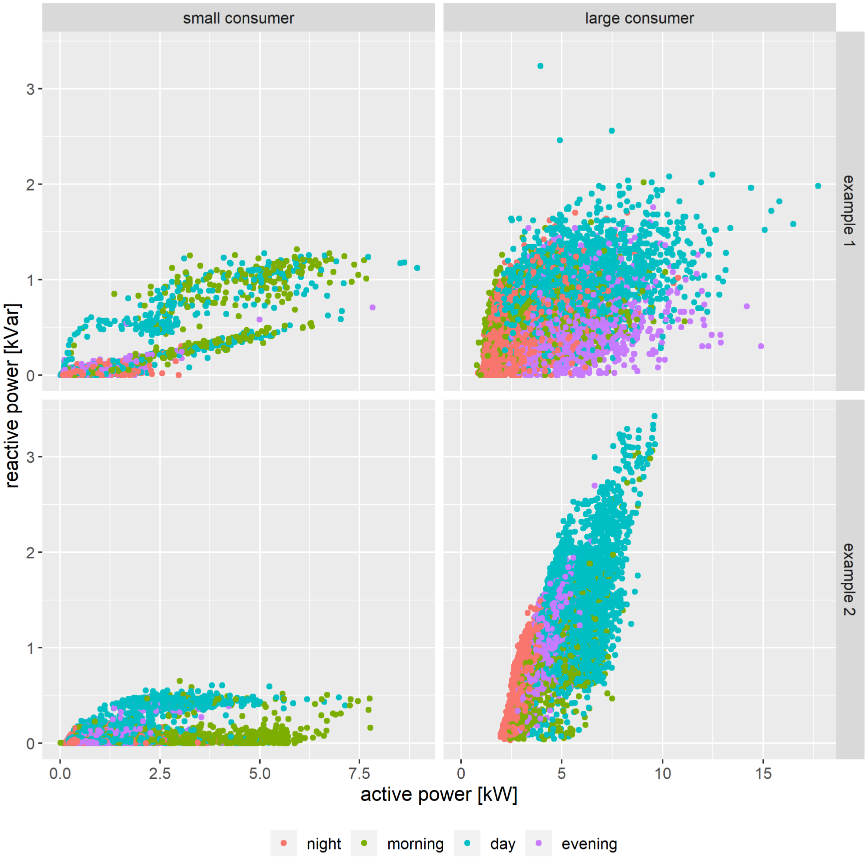


FIGURE 6.2: Scatter plots of reactive power with respect to active power for four different consumers from the Costa Rican smart meter data set.

still provide some clues to estimate the power factor value. More precisely, it comes out that the size of the consumer, the period of the day, and the value of active power are major determinants. Indeed, smaller consumers are probably residential end-users who do not possess the same appliances as larger consumers like commercial end-users. Besides, different electrical appliances with various power factors are used over the course of the day. Furthermore, the active power value at a certain time step gives an indication of the appliances in use and, consequently, of the corresponding reactive power consumption.

Appliance	Power factor
Washing machines	0.55-0.59
Clothes dryer (with heating)	1
Clothes dryer (w/o heating)	0.44-0.47
Central air conditioner	0.90-0.92
Central air conditioner (fan only)	0.54-0.56
Water heater	1
Electric range / oven	0.95-0.98
Dishwasher (pre-wash)	0.62-0.65
Dishwasher (wash and dry cycles)	1
Refrigerator (normal mode)	0.989-0.999
Refrigerator (defrost cycle)	1

TABLE 6.1: Power factor of standard domestic appliances as reported in [287].

The suggested synthesis approach leverages these different observations and assumptions. Based on a training set of smart meter data, the power factor is first calculated from active and reactive power measurements for each consumer and each time step:

$$pf_{i,t} = \frac{P_{i,t}}{\sqrt{P_{i,t}^2 + Q_{i,t}^2}}, \quad \forall i \in \{1, 2, \dots, n_{\mathrm{SM}}\},\ \forall t \in \{1, 2, \dots, T\}, \tag{6.12}$$

where $pf_{i,t}$ is the power factor of consumer i at time t, and $P_{i,t}$ and $Q_{i,t}$ are the active and reactive power values recorded at consumer i at time t. In addition, n_{SM} is the number of smart metered consumers (i.e., with active and reactive power measurements), and T is the total number of time steps under consideration. Subsequently, all calculated power factor values are categorized according to each combination of the following determinants:

- Consumer size: Training consumers are clustered by the k-means algorithm according to their average power consumption. The number of clusters is generally low and depends on the variety of consumers.

- Period of the day: The day is split into four main periods, i.e., morning (5:00-9:00), day (9:00-17:00), evening (17:00-21:00), and night (21:00-5:00)

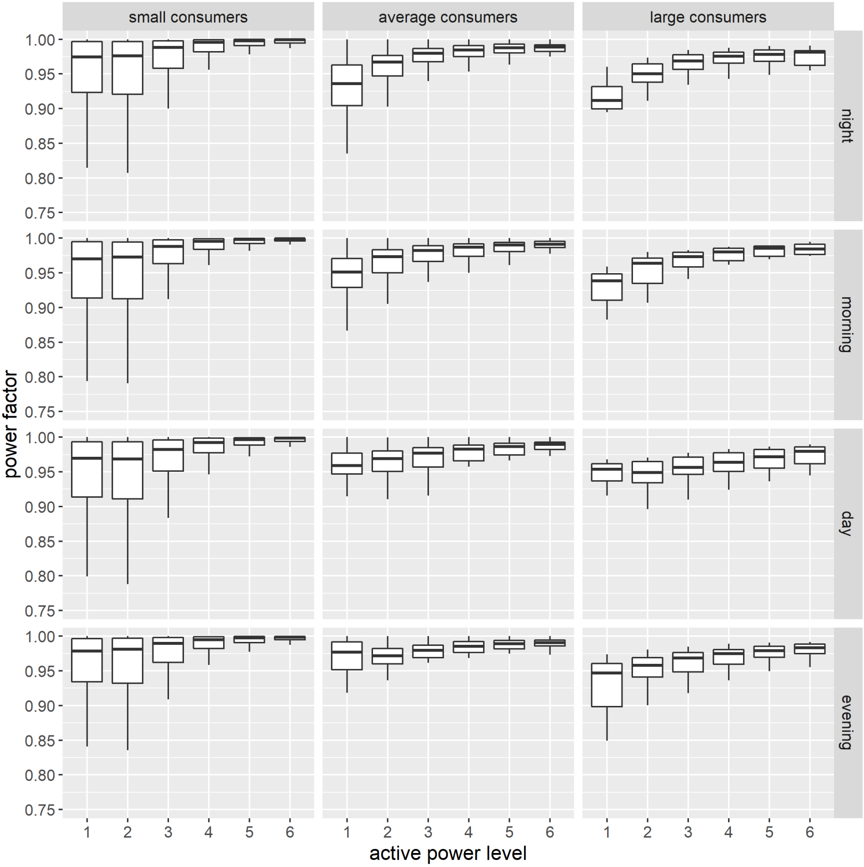


Figure 6.3: Box and whisker plots of the power factor of 1’000 Costa Rican consumers categorized by the consumer size, period of the day, and active power level. Outliers are not shown for the sake of clarity.

- Active power level: For each training consumer, active power values are clustered by the *k*-means algorithm into several power levels. The number of levels results from a trade-off between accuracy and robustness of the outcome.

The variation of the power factor according to the combination of these determinants is illustrated in Figure 6.3. The average values per combination are summarized in Table 6.2. In this case, power measurement data come from a random selection of 1’000 consumers in the Costa Rican smart meter

Intra-day period	Consumer size	Average power factor (from active power level 1 to 6)
Night	small	0.920 | 0.921 | 0.960 | 0.977 | 0.985 | 0.991
Night	average	0.934 | 0.958 | 0.970 | 0.976 | 0.980 | 0.982
Night	large	0.910 | 0.943 | 0.962 | 0.970 | 0.974 | 0.975
Morning	small	0.923 | 0.920 | 0.962 | 0.979 | 0.986 | 0.991
Morning	average	0.945 | 0.965 | 0.973 | 0.977 | 0.980 | 0.984
Morning	large	0.925 | 0.953 | 0.966 | 0.973 | 0.979 | 0.980
Day	small	0.921 | 0.923 | 0.955 | 0.974 | 0.982 | 0.988
Day	average	0.952 | 0.961 | 0.967 | 0.972 | 0.976 | 0.980
Day	large	0.939 | 0.944 | 0.954 | 0.961 | 0.967 | 0.974
Evening	small	0.940 | 0.945 | 0.965 | 0.978 | 0.985 | 0.989
Evening	average	0.971 | 0.965 | 0.972 | 0.977 | 0.981 | 0.984
Evening	large	0.923 | 0.946 | 0.961 | 0.969 | 0.975 | 0.979

TABLE 6.2: Average values of the power factors illustrated in Figure 6.3, categorized by intra-day period, consumer size, and active power level.

data set[3]. All consumers are clustered into three groups, and six active power levels are selected. First of all, a high variance can be seen for small consumers and at lower active power levels. This is mainly explained by the fact that a majority of data points belong to these categories. In addition, most appliances consume at lower power levels but do not necessarily exhibit the same power factor. This complicates its estimation in the absence of actual sub-metering data. Nevertheless, it is still visible from smart meter data that the power factor tends to increase at higher active power values. This is in line with the flattening of reactive power values with respect to active power, as observed in Figure 6.2. Moreover, the larger the consumer, the lower the power factor, which is also visible in Figure 6.2. Besides, the power factor of average and large consumers at the first active power level drops during the night with respect to the day. For small and average consumers, an increase is visible in the evening at lower power levels with respect to other periods of the day. Overall, although the difference between the categories might seem minor, it must be noted that a small variation of the power factor with

---

[3] Data preparation is presented in 5.2.2.

fixed active power can still lead to a significant change in reactive power, as expressed by Equation (6.9).

Finally, the average values for each combination of the above-mentioned determinants form a look-up table, as given by Table 6.2, from which the power factor for a consumer without metered reactive power can be selected:

$$pf_l = \underset{k \in \Lambda}{\text{mean}} \left(pf_k \right), \quad \forall l \in \Lambda, \tag{6.13}$$

where pf_l and pf_k are the power factors allocated to consumer l and measured at the smart metered consumer k, respectively. The power factor values considered in the averaging process must comply with the conditions of Λ. The conditions of Λ are the features of consumer l at a certain time instant, namely a combination of a consumer size, a period of the day, and an active power level.

6.2.4 *Optimal Aggregate Matching*

This section details the selection and assignment of load profiles to actual non-metered end-users in a distribution grid. On the one hand, synthetic active and reactive power profiles must fill the resulting power gap between the aggregation of all smart metered consumer's profiles and the available power measurements at an aggregate level (e.g., at the feeding transformer), as defined in Equation (6.1). As a necessary condition, smart meter measurements must be included in the aggregate measurements. On the other hand, any piece of information from non-metered end-users can be leveraged to customize the synthetic profiles. Normally, the monthly or yearly energy consumption is available for billing purposes. Alternatively, this information can be statistically estimated for each consumer. Consequently, the suggested approach ensures optimal aggregate matching in both spatial and temporal dimensions. This innovative approach has first been presented in [288] before being adapted in [308]. Note that the standard load allocation presented in Section 6.2.2.1 does not require this step since it is directly designed for perfect spatial and temporal matching.

6.2.4.1 *Matching with Aggregate Power Measurements*

In a first stage, a set of load profiles are selected to optimally match with the active power measurements at an aggregate level. For that purpose, at least twice as many synthetic profiles as the number of non-metered consumers are created based on the available smart meter data. According to a preliminary sensitivity analysis, this number ensures a sufficiently large pool of synthetic load profiles. Subsequently, following binary optimization problem selects the most appropriate profiles to fill the active power gap:

$$\min \beta \quad |P_{\text{gap}} - P_{\text{synth}} \cdot \beta| \tag{6.14a}$$

$$\text{s.t.} \quad \sum_{j=1}^{n_{\text{synth}}} \beta_j \geqslant \alpha \cdot n_0, \tag{6.14b}$$

where $P_{\text{gap}} \in \mathbb{R}^{T}_{\geqslant 0}$ is the power gap profile, $P_{\text{synth}} \in \mathbb{R}^{T \cdot n_{\text{synth}}}_{\geqslant 0}$ is a matrix of synthetic profiles, and $\beta \in \{0,1\}^{n_{\text{synth}}}$ determines whether a synthetic profile is selected. In addition, T is the number of considered time steps, n_{synth} is the total number of synthetic profiles, and n_0 is the number of non-metered consumers. Finally, α is a scaling parameter associated to n_0, which allows Constraint 6.14b to guarantee a minimum number of selected profiles (i.e., associated with $\beta_j = 1$).

Subsequently, reactive power is directly synthesized based on the selected active power profiles. In order to match with the reactive power gap at an aggregate level, individual reactive power profiles are scaled as follows:

$$Q'_{\text{synth},i} = \frac{Q_{\text{gap}}}{\sum_{j=1}^{m} Q_{\text{synth},j}} \cdot Q_{\text{synth},i}, \quad \forall i \in \{1, 2, \dots, m\}, \tag{6.15}$$

where $Q'_{\text{synth},i}$ and $Q_{\text{synth},i}$ are the scaled and original synthetic reactive power profiles, respectively. In addition, Q_{gap} is the reactive power gap profile as defined in Equation (6.2), and m is the number of selected synthetic load profiles. In contrast to active power profiles, synthetic reactive power profiles undergo a transformation of their shape.

6.2.4.2 *Matching with Individual Energy Requirements*

In a second stage, selected synthetic profiles are assigned to the non-metered consumers such that the mismatch between their reported energy consumption and the resulting energy consumption given by the synthetic profiles is minimized. Since generally only active power consumption is considered in billing data, the load assignment step is solely based on active power data. Per definition, the sum of the energy consumption values of all non-metered consumers must be close enough to the energy resulting from the power gap profile. A small mismatch might still be explained by an inaccurate estimation of active power losses, by the error margin of smart meters, or by the presence of loads in the system which are not reported (e.g., non-technical losses).

Formally, the optimal allocation of m synthetic load profiles to n_0 non-metered consumers is based on the principle of the bin packing problem [309]. This is an iterative problem which can be represented as the successive packing of m items of various sizes (i.e., energy consumption of synthetic loads) into n_0 different containers (i.e., energy requirement of non-metered consumers). This problem can be solved by means of a greedy algorithm. Beforehand, the energy consumption values of each selected synthetic profile are sorted in descending order (i.e., $E'_1 \geqslant \ldots \geqslant E'_m$). This defines the order in which synthetic load profiles are allocated. Besides, the (remaining) energy requirement for each non-metered consumer is initialized at the reported energy requirement value:

$$E_i^{(0)} = E_i, \quad \forall i \in \{1, 2, \ldots, n_0\}, \tag{6.16}$$

where $E_i^{(0)}$ is the energy to be allocated to non-metered consumer i at the initialization step, and E_i is the total energy requirement of non-metered consumer i. Subsequently, each iteration step $j \in \{1, 2, \ldots, m\}$ assigns the j^{th} largest synthetic load to the non-metered consumer which will minimize the Euclidean norm of the vector of remaining energy requirements:

$$i = \underset{k}{\operatorname{argmin}} |(E_1^{(j)}, \ldots, E_k^{(j)}, \ldots, E_{n_0}^{(j)})|_2, \quad \text{with } E_k^{(j)} = E_k^{(j-1)} - E'_j \tag{6.17}$$

where i is the index of the consumer who gets assigned with the j^{th} largest synthetic load profile, $E_k^{(j)}$ is the remaining energy requirement of consumer k at iteration step j, and E'_j is the energy consumption value of the j^{th} largest synthetic load.

Assuming that the number of synthetic load profiles m is approximately equal to the number of non-metered consumers n_0, one single synthetic load profile is generally assigned to each non-metered consumer. Nevertheless, the smallest consumers might not receive a profile, whereas the load of large consumers might consist of the spatial aggregation of multiple profiles. Although a small portion of non-metered consumers might be modeled with a zero power consumption profile, this is acceptable since their actual impact on the system is anyway limited. A constraint to ensure the assignment of at least one profile per consumer could still be added, but at the cost of a generally sub-optimal load assignment. Finally, note that the load profile selection defined in Equation (6.14) and the bin-packing problem are under-constrained problems that can have multiple feasible solutions.

6.3 CASE STUDIES

Based on different case studies, this section analyzes the performance of the various approaches presented in Section 6.2. First, active power pseudo-measurements are synthesized and evaluated in a sub-grid of the City of Basel. In this case, active power measurements of smart metered end-users and at the transformer level are available, together with the billing data of all end-users. Second, the synthesis of reactive power profiles relies on Costa Rican smart meter data, where both active and reactive power measurements are recorded.

In order to quantitatively evaluate the performance of the different models, following metrics are considered:

$$\text{MAPE} = 100\% \cdot \frac{1}{|\Omega|} \cdot \sum_{j \in \Omega} |\frac{y_{\text{synth},j} - y_{\text{obs},j}}{y_{\text{obs},j}}|, \tag{6.18a}$$

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega|} \cdot \sum_{j \in \Omega} (y_{\text{synth},j} - y_{\text{obs},j})^2}, \tag{6.18b}$$

where $y_{\text{obs},j}$ and $y_{\text{obs},j}$ are the observed and synthetic values of element j (e.g., time step j, consumer j), respectively. In addition, Ω is the set of all elements, and $|\Omega|$ refers to the total number of elements. In both cases, the lower the value, the better the accuracy.

6.3.1 *Active Power Pseudo-Measurements*

The performance of the three active power synthesis methods and of the optimal aggregate matching approach is evaluated in the residential neighborhood of the City of Basel illustrated in Figure 3.5. As a reminder, the grid consisted of 6 metered PV systems and 583 end-users, of which 320 end-users (i.e.,55%) were equipped with a reliable smart meter at the time of data preparation. For the purpose of this study, a time period of one year is considered. Measurement data are characterized by a temporal resolution of 15 minutes. Data preparation is performed as described in Section 4.3.

Figure 6.4 illustrates the power gap in a summer day and a winter day between the aggregated profiles of metered consumers and PV systems and the transformer loading. Note the considerably higher consumption in winter, especially in the evening. In addition, the transformer experiences reverse power flows around noon, which is potentially problematic if its protection system has not been designed accordingly. Regarding the power gap, its

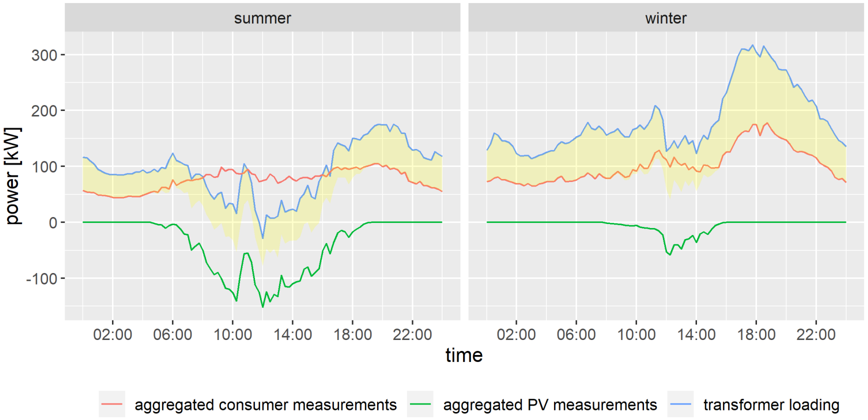


Figure 6.4: Active power gap (i.e., yellow shape) over a typical summer and a winter day between the aggregated smart metered profiles and the transformer loading in the residential neighborhood presented in Figure 3.5.

shape is similar to the aggregation of the smart metered consumer profiles. This indicates that the load profiles of non-metered consumers have similar characteristics as the metered consumer profiles at a spatial aggregate level. Since all consumers are located in the same neighborhood, it is also reasonable to assume that individual metered and non-metered consumers have similar statistical properties. Hence, the MCM-based synthesis approaches leverage smart meter data from the same neighborhood in order to create pseudo-measurements for non-metered end-users.

In the following, the proposed study first focuses on the properties of individual synthetic load profiles. In this case, the MCM-based load profiles are directly compared with the smart meter data measurements on which they are based. According to a preliminary sensitivity analysis, the number of states and sub-levels does not significantly influence the outcome as long as they lie above a minimum threshold. As a trade-off between computational complexity and accuracy at both individual and aggregate levels, the number of states is set to 6 for the two presented MCM-based approaches, and ten sub-levels are selected for the traditional MCM. Moreover, the benchmark SLA relies on the aggregation of all metered consumers. In the second stage, synthetic load profiles are allocated to actual end-users, and the performance at a spatial and temporal aggregate level is considered. Note that SLA-based

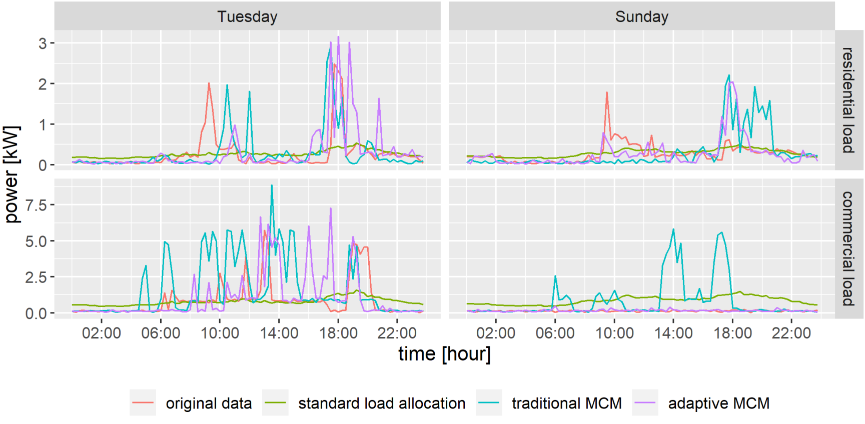


FIGURE 6.5: Illustration of metered and synthetic active power profiles on two distinctive weekdays for a residential and a commercial consumer.

synthetic load profiles are only evaluated at the individual consumer level. Per definition, they are perfectly consistent at an aggregate level.

6.3.1.1 *Properties at the Consumer Level*

First of all, Figure 6.5 allows for qualitative comparison between the actual active power measurements of a residential and a commercial consumer and the respective synthetic load profiles. For comparison purposes, the MCM-based approaches are directly trained on the respective smart meter data. It appears that the standard load allocation creates very flat and smooth profiles which fail to reproduce the high load volatility. Conversely, both MCM-based approaches are able to generate active power spikes of similar magnitude as the observations. In addition, the adaptive MCM is able to predict the absence of activity on Sunday for the commercial load, in contrast to the traditional version. Overall, active power pseudo-measurements can obviously not faithfully represent the load of actual consumers at each time step. This does not depend on the specific synthesis approach but on the information provided by the AMI system (see the introduction of Section 6.2), which does not allow for such accurate representation. Nevertheless, pseudo-measurement synthesis approaches must still be able to generate load profiles that are visually realistic and statistically reflect the load distribution of actual consumers. Accordingly, a point-wise comparison of synthetic with metered load profiles would not make sense. Volatile synthetic data would

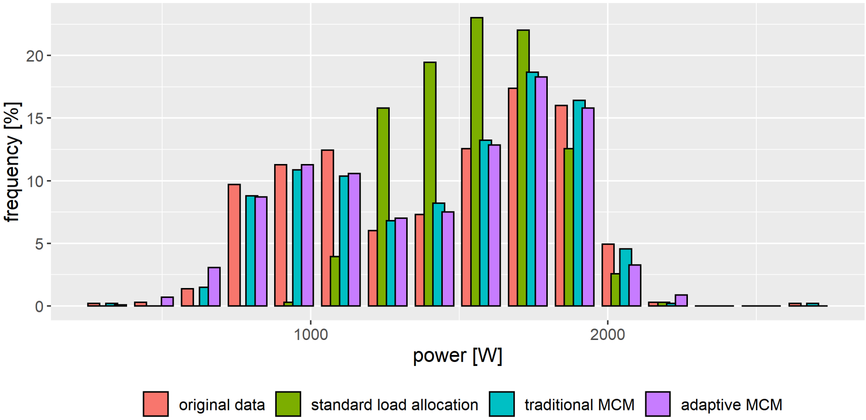


Figure 6.6: Load distribution of the original and synthetic profiles for a specific large consumer between 18:00 and 19:00.

be prejudiced in comparison with smooth data due to the so-called double penalty effect, as detailed in Section 9.3. In contrast, this section focuses on the faithful representation of the statistical properties of individual load profiles.

A realistic distribution of power values is one of the properties that pseudo-measurements must be able to reflect. This is illustrated in Figure 6.6 for a large consumer. In this case, only data between 18:00 and 19:00 are considered with the aim of assessing the ability of algorithms to adapt to variations over the day. It appears that the standard load allocation basically generates data with a Gaussian distribution, which is far from the original distribution at this specific hour. Conversely, both the traditional and adaptive MCM approaches reproduce almost perfectly the original load distribution.

Moreover, Figure 6.7 evaluates the ability of the synthesis approaches to account for seasonality on different time scales. More precisely, each data point corresponds to the mismatch in energy for a specific end-user in predefined temporal conditions:

$$\Delta E_{i,k} = 100\% \cdot \frac{E_{\mathrm{synth},i,k} - E_{\mathrm{SM},i,k}}{E_{\mathrm{SM},i,k}}, \quad \forall i \in \{1, 2, \ldots, n_{\mathrm{SM}}\},\ \forall k \in \Gamma, \tag{6.19}$$

where $\Delta E_{i,k}$ is the relative mismatch in energy of consumer i in temporal conditions k. $E_{\mathrm{synth},i,k}$ and $E_{\mathrm{SM},i,k}$ are the total energy consumption values visible in the smart metered and synthetic load profiles of consumer i in

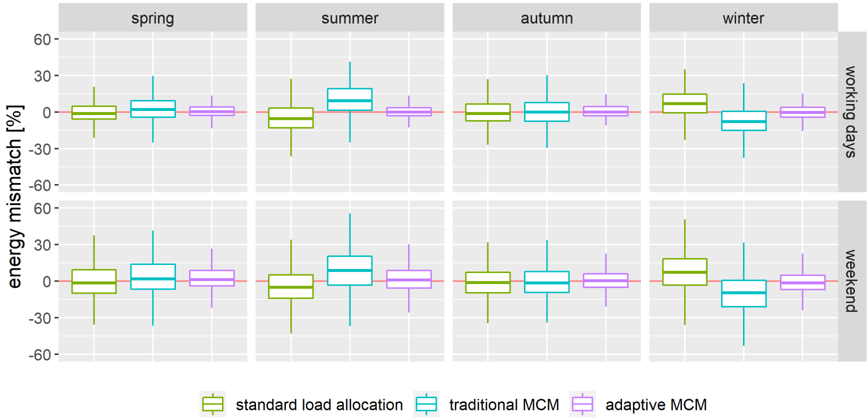


FIGURE 6.7: Box and whisker plots of the mismatch in energy between metered and synthetic data, evaluated at different seasons and types of weekday.

temporal conditions k, respectively. In addition, n_{SM} is the number of smart metered consumers, and Γ is the set of all possible combinations of a season with a type of weekday (i.e., working day or weekend). The AMCM exhibits an average mismatch in energy very close to zero and a relatively low variance for all seasons and types of weekday. In contrast, the energy consumption is generally overestimated by the traditional model in summer (by 26% on average) and slightly underestimated in winter (by 7% on average). This comes from the fact that its training phase does not account for the seasons, although they influence the energy consumption (i.e., higher consumption in winter than in summer). To a lesser extent, the opposite tendency is visible for the standard load allocation. The seasonality effect on energy consumption appears to be stronger at an aggregate level than for the average individual end-user. Finally, the weekend is characterized by a higher variance than working days, but the general trends remain unchanged.

6.3.1.2 *Properties at an Aggregate Level*

The selection of the most suitable synthetic profiles is performed according to Equation (6.14). The binary optimization problem is implemented and solved in MATLAB with the help of the optimization toolbox YALMIP [310] and the solver Gurobi [311] which can deal with mixed-integer conic problems. As input for the optimization problem, 640 yearly load profiles are synthesized

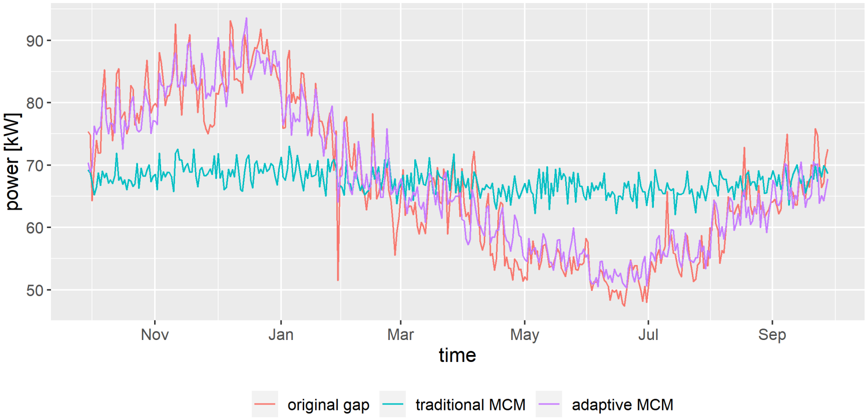


Figure 6.8: Evolution over one year of the average power consumption of the original gap profile and of the aggregation of optimally selected synthetic profiles.

based on the active power measurements of the 320 smart metered consumers. In other words, one Markov chain model is created for each metered consumer, which serves for the synthesis of two load profiles. Based on a temporal resolution of 15 minutes, each yearly profile consists of $4 \cdot 24 \cdot 365 = 25'440$ data points. Unfortunately, this amount of data points per profile is not computationally tractable for the binary optimization problem within a reasonable time. In order to lower the problem complexity, the optimization problem only considers the average power consumption per day. Hence, matrix P_{synth} in Equation (6.14) is reduced to a dimension of 365×640. Moreover, the scaling parameter α is set to 1 in order to ensure the selection of at least one load profile per non-metered consumer on average. In fact, this constraint is binding for both types of synthetic input profiles, meaning that the same number of synthetic profiles as the number of non-metered consumers are selected. Note that the properties of the synthetic active power profiles are not affected by the selection process.

Figure 6.8 illustrates the yearly profiles of the original active power gap and of the aggregation of optimally selected synthetic profiles[4]. For the sake of clarity, power values are averaged over each day. On average, the power of the gap profile approximately varies from 90 kW in December to

---

[4] As previously mentioned, the standard load allocation is already designed for a perfect aggregate matching and is therefore not considered here.

50 kW in June. The aggregation of AMCM-based profiles nicely reflects this yearly seasonality, which is however not the case for the traditional version. This is in line with the characteristics of the respective individual synthetic profiles visible in Figure 6.7. These observations are also confirmed by the MAPE between the original gap and the aggregations of synthetic profiles, which amounts to 24.1% and 19.2% for the traditional and adaptive MCM, respectively[5]. Finally, the optimally selected synthetic profiles are allocated to the non-metered consumers via the bin packing problem. It appears that the five smallest non-metered consumers are not assigned with a synthetic load profile, whereas five other consumers eventually consist of the aggregation of two synthetic profiles. In terms of a mismatch between the reported yearly energy and the assigned energy, the traditional MCM and adaptive MCM lead to a MAPE of 7.95% and 8.46%, respectively. On this aspect, the traditional MCM slightly outperforms the adaptive version.

6.3.2 *Reactive Power Pseudo-Measurements*

Since smart meters in the City of Basel do not record reactive power, the evaluation of reactive power synthesis relies on the Costa Rican smart meter data set, where both active and reactive power quantities are recorded. For that purpose, the three presented synthesis approaches are tested on 1’000 randomly chosen consumers over a period of four months. The adaptive power factors are based on the look-up table presented in Table 6.2. It must nevertheless be noted that the training consumers (i.e., used for the creation of the look-up table) are different from the test consumers (i.e., used in the evaluation). The average and aggregate power factors are calculated based on the spatial aggregation of both the training and test consumers. This represents the situation of a large distribution grid with 50% smart meter penetration[6]. In this case study, the average power factor at the aggregate level is equal to 0.967 inductive.

Figure 6.9 illustrates the synthesis of reactive power over one day for one small and one large consumer. For comparison purposes, power values are

---

[5] When considering only the daily average power values, the MAPE decreases to 14.4% and 4% for the traditional and adaptive MCM, respectively.

[6] In a real grid, note that a small portion of active and reactive power visible at an aggregate level is consumed (or produced) by power lines and transformers. In addition, PV systems usually produce purely active power but are sometimes required to generate or consume reactive power for grid support purposes. Hence, power factor values inferred at a transformer or substation level might slightly vary from the power factor values calculated over the aggregation of all metered consumers.

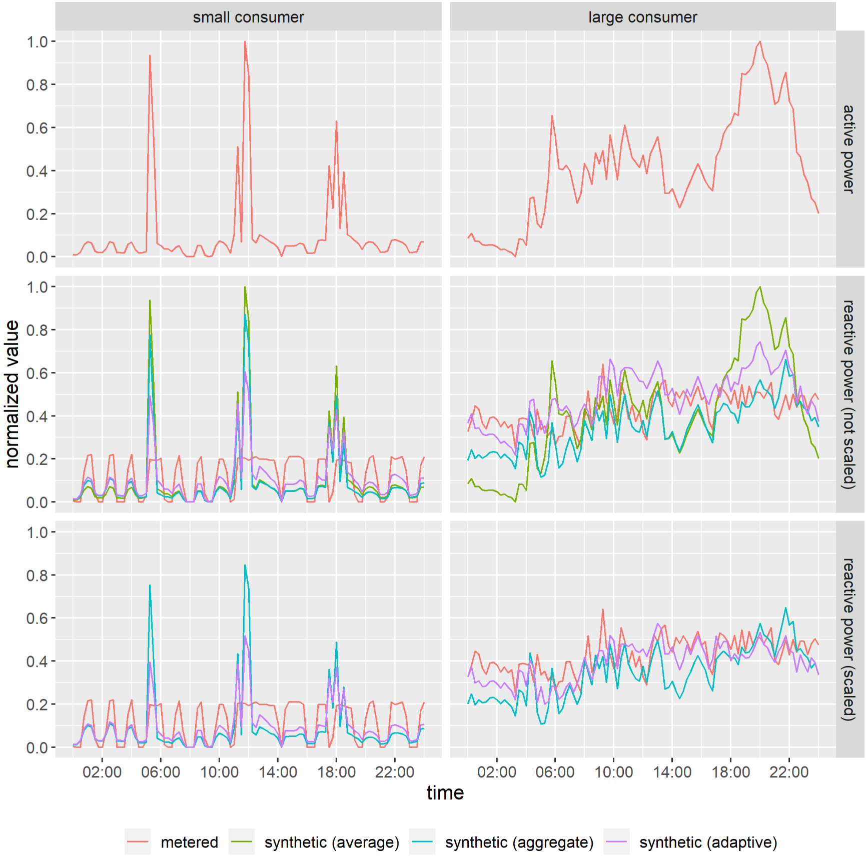


FIGURE 6.9: Illustration of reactive power synthesis over one day for a small and a large consumer, with and without scaling (i.e., spatial aggregate matching). For better visualization, active and reactive power values are normalized between 0 and 1.

normalized between 0 and 1. Regarding the small consumer, active power measurements slightly oscillate at a low level over most of the day and exhibit high values during three short periods. Only the oscillating pattern appears in reactive power measurements. Hence, high power activities certainly belong to purely resistive appliances. To different extents, the shape of the active power profile is translated into reactive power pseudo-measurements by all synthesis techniques. The average and aggregate power factors lead to a faithful reproduction of the shape of active power measurements. In contrast, the adaptive version adjusts the power factor according to the conditions such that the resulting reactive power profile is closer to the metered data. In any case, reactive power is overestimated during active power spikes and underestimated at low active power levels. In other words, the power factor is underestimated by all synthesis approaches for high power levels and overestimated for low power levels[7]. Besides, the spatial aggregate matching (or scaling) does not particularly affect the small consumer. Still note that the scaled versions of the reactive power profile created by the average and aggregate power factors are identical. Regarding the large consumer, reactive power measurements do not specifically follow the pattern of active power. Without scaling, the average power factor fails to properly estimate reactive power, which is underestimated at night and overestimated in the evening. In contrast, both aggregate and adaptive versions adapt their power factor over the day, which apparently leads to more realistic reactive power pseudo-measurements. Nevertheless, the scaling process seems to improve the correspondence of all reactive power profiles, which do not significantly differ from each other anymore.

Furthermore, the RMSE is calculated for each consumer between the actual reactive power measurements and the synthetic profiles. The outcome is displayed in Figure 6.10, where each data point corresponds to the error of a certain synthesis approach for a specific consumer. Without scaling, it appears that the average version is outperformed by the competing methods. For example, the median error drops by 12.1%, 17.8% with respect to the use of aggregate and adaptive power factors, respectively. The respective median drop even amounts to 42.4%, 14.8% for large consumers. These data also show that the adaptive version is slightly more appropriate for small consumers, whereas the aggregate version is clearly beneficial for large consumers. Nevertheless, spatial aggregate matching equalizes and increases the estimation accuracy of all approaches, especially for large consumers. By definition, the average and aggregate power factors lead to the exact same reactive power profiles after scaling. In this case, the adaptive version still outperforms the competing versions for small consumers (i.e., the median RMSE drops by 15.3%). The advantage is substantially reduced for large consumers (i.e., the median RMSE drops by 2.7%).

---

[7] It must be kept in mind that power factor values come from the look-up table and result from an average over all consumers in the training data set. Hence, for other small consumers, the power factor is probably overestimated for high power values and underestimated for low power values.

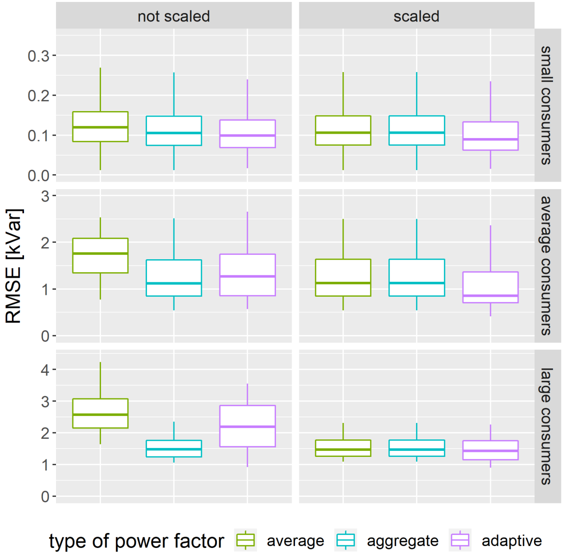


FIGURE 6.10: Box and whisker plots of the root-mean-square error between metered and synthetic reactive power profiles, categorized by synthesis approach and consumer size. Outliers are not shown for the sake of clarity.

Although the spatial aggregate matching allows for better performance, the shape of reactive power pseudo-measurements is relatively similar to the respective active power profiles, especially for small consumers. This still leads to substantial absolute errors. Concretely, the assumption of a time-invariant power factor is clearly not pertinent, but its approximation based on a look-up table or based on observations at an aggregate level is also not fully satisfying. In fact, actual reactive power cannot be properly approximated by an average behavior over a large set of examples. A more accurate point-wise estimation of reactive power could be obtained by ML-based algorithms which potentially better adapt to the characteristics of individual consumers. Reactive power synthesis approaches could also benefit from load detection or load disaggregation techniques, as presented in Chapter 8, which give insight into the type of some appliances in usage based on active power measurements.

6.4 CONCLUSION

To conclude, a comprehensive procedure to synthesize both active and reactive load profiles and allocate them to individual non-metered consumers is described in this chapter. Particular focus is given to the realistic representation of synthetic load profiles without excessive effort. Presented synthesis approaches are only based on traditionally available information in distribution grids. They rely on smart meter data and aggregate measurements but do not necessitate sub-metering data or specific customers' information other than their respective average energy consumption. The chapter also specifically addresses the compliance of individual synthetic data with aggregate information in a given distribution grid. The proposed procedure and approaches are tested in real-world residential areas which mainly consist of households and a few commercial end-users. Adaptations might be needed for other types of consumers, especially for industrial loads.

In terms of contribution, an adaptive Markov chain model is first proposed for active power synthesis. Markov chain models are stochastic algorithms inherently designed to reflect the volatility and value distribution of the input data. By representing transition probabilities by a logistic regression model, the adaptive version can additionally grasp the notion of seasonality or periodicity at different time scales. This is not practicable by the traditional version without an extensive amount of training data and the hard-coding of the notion of seasonality. The proposed approach is also compared to the standard load allocation. This simple method is often leveraged by DSOs and in the literature but totally fails to reflect the statistical properties observed in smart meter data. Further advanced pseudo-measurement synthesis approaches proposed in the literature (e.g., autoregressive models, clustering algorithms, ANNs, GBTs) are not explicitly considered in the performance evaluation. As detailed in Section 9.3, they nevertheless lead by design to a certain smoothing of the resulting load profile and lose part of their statistical properties, in a similar way as the standard load allocation. Mainly designed for the synthesis of pseudo-measurements, the adaptive Markov chain model can also lend itself to the imputation of missing data and to forecasting highly volatile loads, as proposed in Sections 3.3.5 and 9.3, respectively. Moreover, Chapter 7 elaborates on the advantages of using AMCM-based synthetic profiles and points out the risks of standard load allocation in the context of DSSE.

Even though they are indispensable in DSSE or power flow simulations, reactive power pseudo-measurements are generally neglected in the literature.

Hence, the chapter also presents a close reflection on this topic via the estimation of power factor values. In fact, reactive power consumption mainly depends on the type of appliances in use, which is barely detectable in most cases solely based on standard smart meter data. Some features such as the consumer size, the period of the day, and the active power level are nevertheless observable and contribute to a rough estimation of the time-variant power factor for a given consumer. A proposed approach leverages these observations from a training smart meter data set in order to build a look-up table with average power factor values in different conditions. Unfortunately, power factors cannot be precisely inferred based on observable information. Especially, average power factor values do not properly represent the large variance observed in actual data. Hence, the proposed look-up table does not provide considerable benefits in comparison with power factor values derived from an aggregate level. In any case, scaling individual reactive power profiles with respect to aggregate measurements contributes to substantially more realistic outcomes. Despite this somewhat mixed performance, the topic definitely merits closer scrutiny. For example, additional features (e.g., weekday, temperature) can serve to fine-tune the power factor estimation. An ML-based prediction algorithm can also probably better infer the relationship between observable features and reactive power than a purely statistical model. Besides, Chapter 8 demonstrates that some specific appliances (e.g., water heater, refrigerator) can still be detected in standard smart meter data. Such knowledge should be leveraged in the synthesis of reactive power pseudo-measurements.

Special attention is finally given to match realistic load profiles both with measurements at a spatial aggregate level and with energy requirements of single non-metered loads. In fact, this aspect is perfectly addressed by the standard load allocation approach, but at the cost of highly unrealistic individual profiles. Besides, the literature sometimes recommends scaling active power pseudo-measurements directly with respect to aggregate measurements, which nevertheless alters the properties of the original load profiles. In this work, the selection and assignment of synthetic load profiles are successfully achieved by solving a binary optimization and a bin packing problem, respectively. Note that this procedure only requires a sufficiently large pool of diverse load profiles with potentially similar statistical properties as the load of non-metered consumers. If this is achievable with actual smart meter data (e.g., from another grid area with similar characteristics), there is no specific need for synthetic load profiles.

7

IMPACT OF DATA ON DISTRIBUTION SYSTEM STATE ESTIMATION

This chapter studies the influence of the AMI design and of the related modeling of pseudo-measurements on the outcome of state estimation in distribution grids, especially at the low-voltage level. A comprehensive sensitivity analysis is carried out that accounts for the type, the penetration level, and the placement of metering devices that compose state-of-the-art AMIs. Special care is also given to the synthesis of power pseudo-measurements, which substantially impacts the estimation of peak values. Although crucial for system operators, peak values are very often neglected in the state estimation literature. For that purpose, the chapter relies on evaluation metrics that are not purely based on the point-wise precision but also consider the statistical properties of the outcome. The evaluation focuses on power injection, bus voltage, and line loading. The sensitivity analysis is performed on an actual 971-bus grid with the corresponding smart meter data from the City of Basel. This chapter is based on [308].

7.1 INTRODUCTION

Supported by the large-scale roll-out of Smart Meters (SMs), the digitalization of distribution grids also reaches the Low-Voltage (LV) level, which was traditionally seen as a black box. Current Advanced Metering Infrastructures (AMIs) provide more and more information at the level of end-users, but also at key points of the Low-Voltage (LV) grid (e.g., cable distribution cabinets, MV/LV transformers). This facilitates an unprecedented variety of new applications such as congestion management, optimal voltage regulation, and quantification of flexible load for demand response purposes [3, 4]. However, the quantity, quality, and type of the measured data streams can greatly vary between different AMIs, which impacts the data-based modeling accuracy of the corresponding distribution grids. Distribution System Operators (DSOs) are confronted with the challenge of developing a cost-effective AMI and properly transforming the relatively new and diverse measurement data into valuable information [114]. Especially, DSOs intend to find a trade-off between the costs of measurement devices and their benefits.

Distribution System State Estimation (DSSE) algorithms allow for the estimation of the most probable state of a distribution grid based on measurement data, which is associated with different challenges. In order to minimize the costs, DSOs normally possess a limited set of metering devices, and the question of their optimal placement arises. For example, the authors in [312–316] present various robust algorithms for the placement of Phasor Measurement Units (PMUs) and/or smart meters, accounting for the meter deployment costs, the uncertainty of distributed generation, and topological reconfigurations. Nevertheless, a large number of metering devices are currently already installed in distribution grids, not based on an optimal placement approach but on immediate needs. Existing studies do not consider the achievable modeling accuracy given a certain sub-optimal meter placement. Besides, the observability of the system is a prerequisite for State Estimation (SE) algorithms but is generally not achievable on the sole basis of the actual measurement data, especially at the LV level. Hence, pseudo-measurements are inevitably required and are traditionally designed to minimize the SE error [296, 297, 317]. By aiming for this goal, pseudo-measurements proposed in the current literature consist of unrealistically smoothed out profiles, whereas more realistic but volatile synthetic data are not an option. This problem is pointed out by the authors in [293] in the case of standard load profiles which provide a biased image of the load distribution in the actual distribution grid. Specifically, peak values must be correctly represented since they define the dimensioning of the distribution grid infrastructure and of distributed resources such as batteries.

Accordingly, this chapter provides comprehensive insights into the SE accuracy that can be achieved depending on the AMI design and on the approach for pseudo-measurement synthesis, with a focus on peak power. Precisely, a sensitivity analysis based on the Weighted Least Squares (WLS) algorithm is carried out in Matlab with respect to the number, the type and the placement of metering devices, and to the approach for pseudo-measurement synthesis. Among others, the analysis compares the influence of active power pseudo-measurements synthesized by the Adaptive Markov Chain Model (AMCM) and the Standard Load Allocation (SLA), thoroughly presented in Chapter 6. Rarely considered in the DSSE literature, the possibility to install measurement devices at distribution cabinets is also included. In total, a benchmark model and 144 different scenarios derived from 6 dimensions with 2 to 3 different options each are studied. It must be noted that many existing studies have separately assessed the influence of pseudo-measurements or the placement of advanced metering devices within the framework of DSSE. However, to the best of the author's knowledge, a comprehensive sensitivity analysis accounting for multiple influencing factors has not been presented so far.

In addition, case studies presented in literature often rely on simple benchmark grids of a few dozen nodes. Solely the Medium-Voltage (MV) level is commonly considered while the LV grid is aggregated at the MV/LV transformer level. Such simplistic models are definitely not representative of actual distribution grids as discussed in Section 3.2. The sensitivity analysis proposed in this chapter leverages the network and measurement data from the City of Basel presented in Figure 3.4. This 971-bus network is representative of a European urban distribution system and notably comprises a local substation feeding multiple MV/LV distribution transformers with their corresponding LV sub-grids. Such configuration is schematically illustrated in Figure 3.2a. The grid contains Photovoltaic (PV) systems as well as residential, commercial, and a few industrial end-users that are connected at LV nodes. The main assumptions concerning the availability of measurement data and the relevant dimensions to consider in the sensitivity analysis are based on discussions with some of the main Swiss and Costa Rican DSOs [55, 56, 223].

The remainder of this chapter is structured as follows. Section 7.2 introduces the Weighted Least Square (WLS) state estimation algorithm and the related challenges in distribution grids, especially in terms of input data. Section 7.3 presents the setup of the sensitivity analysis. The section starts by specifying the network and corresponding measurements used in the case study. Next, a complete description of the six dimensions and the associated options is given, and the evaluation metrics are defined. In this work, the SE performance is not automatically assessed via the traditionally used point-wise metrics (e.g., Root-Mean-Square Error (RMSE)). Especially, an adjusted version of the RMSE is used to mitigate the so-called double penalty effect which prejudices volatile data. Special focus is also given to the proper estimation of peak values. Subsequently, Section 7.4 displays and explains the results of the sensitivity analysis. In addition to bus voltages, the accuracy of active and reactive power bus injections and line loadings are considered. Moreover, the computational complexity and potential scalability of the analysis are discussed in Section 7.5. Finally, Section 7.6 summarizes the principal outcomes and outlines potential future work.

7.2 DISTRIBUTION SYSTEM STATE ESTIMATION

State estimation is a data-driven technique that combines different sources of redundant measurements to find the most probable state of a system. Observability issues require DSSE algorithms to be able to deal with a low redundancy of measurements and to integrate a substantial amount of pseudo-measurements [8, 114]. Although other techniques might be more suitable for distribution grids, this study relies on the well-known WLS algorithm due to its low computational cost, its ability to weight the measurements according to their type, and its extensive use in the SE literature. In the following, the DSSE algorithm and the different types of input measurements are defined.

7.2.1 *Weighted Least Squares Algorithm*

The optimization problem of the WLS algorithm can be formulated as follows:

$$\hat{x} = \underset{x}{\operatorname{argmin}} \left(z - h(x)\right)^T W \left(z - h(x)\right), \tag{7.1}$$

where x and $\hat{x}$ are the true and estimated state vectors, respectively. In addition, z is the measurement vector, h is a measurement function linking the measurements to the states, $W = diag\{\sigma_1^{-2}, \ldots, \sigma_m^{-2}\}$ is a weight matrix, σ_i is the standard deviation of the i^{th} measurement, and m is the total number of measurements. The WLS algorithm assumes that the elements of the error vector $e = z - h(x)$ have a Gaussian distribution and are statistically independent. Hence, the weight associated to a certain measurement is chosen as inversely proportional to the square of its standard deviation. Extensive literature is available for more information on the WLS algorithm, e.g., in [318].

7.2.2 *Measurements and Standard Deviation*

The measurement vector z of the WLS algorithm basically consists of direct measurements, pseudo-measurements, and virtual measurements. The latter two types increase the system observability and prevent the problem from becoming ill-conditioned. Table 7.1 summarizes the types of measurement data that can be typically found in a distribution grid down to the LV level. For each type of measurement data, the corresponding standard deviation value has been tuned with respect to a grid-specific analysis, as recommended by the authors in [319]. As previously mentioned, they define the weight given to each measurement in the DSSE problem. Note that the standard deviation

Quantity	Category	Location	Standard deviation
Power injection	direct	bus (full SM coverage)	*0.01*
Power injection	pseudo	bus (zero SM coverage)	*1*
Power injection	pseudo	bus (partial SM coverage)	*see Eq. 7.2*
Power injection	virtual	“no-load” bus	0.001
Power flow	direct	feeder/transformer/cabinet	*0.01*
Power flow	virtual	“no-load” line	0.001
Power flow	virtual	“adjacent” line	*0.01*
Voltage	direct	feeder/transformer/cabinet	*0.001*
Voltage	direct	bus (with at least 1 SM)	*0.001*
Voltage	pseudo	bus (without SM)	*0.1*

TABLE 7.1: Types of measurements in a distribution grid. Standard deviation values in italic are given in relative terms.

values given in italic are in relative terms (i.e., they must be multiplied with the corresponding measurement value in the per-unit system), whereas the standard deviation values of “no-load” buses and lines are absolute values. The following subsections provide more information on each type of measurement data.

7.2.2.1 *Direct Measurements*

Direct measurements refer to any actual measurements coming from the AMI, either via smart meters at the end-user side or via advanced metering devices at the feeder[1], transformers, and cable distribution cabinets. More precisely, smart meters can measure power injections and voltages, whereas advanced metering devices measure power flows and voltages at the feeder, transformers, and cable distribution cabinets. All measurements are converted to the same time resolution. The standard deviation for direct measurements is set to the expected level of noise indicated in the respective data sheets of the measurement devices. It can typically reach up to 0.1% and 1% of the measured value for voltage and power measurements, respectively. According

---

[1] In this work, the feeder is referred to as the power line feeding the entire system under consideration. It is generally connected to a local substation.

to a preliminary study, relative weights lead to significantly better accuracy than absolute weights for bus voltages and reactive power injections. Besides, the impact of the chosen values for all weights on the estimation of all quantities is negligible as long as they stay in the magnitude of 0.1% and 1% for voltage and power measurements, respectively.

7.2.2.2 *Pseudo-Measurements*

Within the framework of DSSE, synthetic active and reactive power profiles are specifically created to cope with the lack of direct measurements. The literature generally assumes that the power at a certain bus is either fully measured or fully synthetic. This is nevertheless not realistic since both metered and non-metered consumers can be located at the same bus of an LV grid. The determination of the associated standard deviation at a bus with partial SM coverage is not straightforward anymore. Hence, if several end-users are located at the same bus, their power is aggregated, and the corresponding standard deviation is proposed to be calculated as follows:

$$\sigma_{\text{inj},b} = \sigma_{\text{pseudo}} - \gamma_b \cdot (\sigma_{\text{pseudo}} - \sigma_{\text{SM}}), \quad \forall b \in \mathcal{B}, \tag{7.2}$$

where $\sigma_{\text{inj},b}$ is the relative standard deviation of the power injection at bus b, and σ_{pseudo} and σ_{SM} are the relative standard deviations of pseudo-measurements and SM power measurements as defined in Table 7.1, respectively. In addition, γ_b is the share of total energy measured by smart meters at bus b, and $\mathcal{B}$ is the set of all loaded buses in the system. Equation (7.2) is justified by Figure 7.1. Based on the measurements of the case study presented in Section 7.3.1, it illustrates the average Normalized Root Mean Square Error (NRMSE) over all active power injections with respect to the share of energy measured by smart meters at each bus. The RMSE is equivalent to the standard deviation and is normalized by the mean power consumption per bus in this case. The error value appears to be close to 100% at buses with pure pseudo-measurements (i.e., the share of energy measured by SMs is equal to zero) and decreases to the level of the SM measurement noise at buses with full SM coverage. The linear approximation shown in Figure 7.1 and corresponding to Equation (7.2) is close to the local polynomial regression. It is therefore chosen to define the standard deviation for buses with partial SM coverage.

Besides, the voltage value at each bus is known to be around 1 pu. This information can be added as pseudo-measurement with a reasonably large standard deviation of 10% whenever a direct voltage measurement is not available.

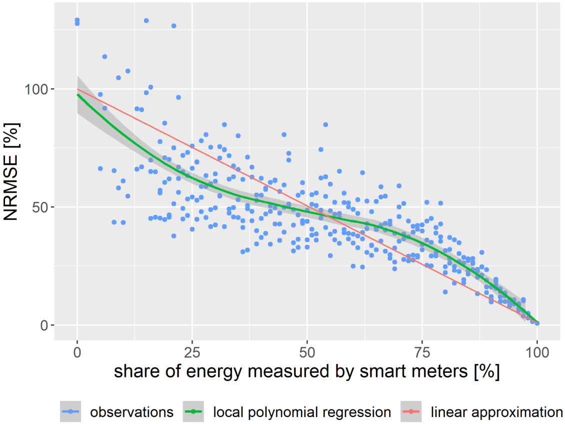


FIGURE 7.1: Average NRMSE over all bus injections in function of the share of yearly energy measured per bus by smart meters.

7.2.2.3 *Virtual Measurements*

Virtual measurements are measurements that are directly derived from the structure of the grid and enable a further increase in the redundancy of measurements. This study first considers zero-power injections and zero-power flows at “no-load” buses and lines, respectively. The information on “no-load” buses generally comes from the metadata. In addition, “no-load” lines are the lines connecting leaf[2] “no-load” buses, which can be inferred from the grid model. These virtual measurements are associated with a very low absolute standard deviation to force the SE algorithm to give an estimate close to zero. Alternatively, “no-load” buses and lines could be modeled as equality constraints in the WLS formulation. This would nevertheless prevent the detection of a misestimation of presumably zero-power quantities.

So-called “adjacent” lines also allow for an expansion of the set of virtual measurements. An adjacent line is defined as a line connected to a metered line via a “no-load” bus. Per definition, its power flow is very similar to the flow of the corresponding metered line. The measurement value and the standard deviation of adjacent lines are set to the same values as for the corresponding metered lines.

---

[2] A leaf bus is a bus connected to only one single line.

7.3 SETUP OF SENSITIVITY ANALYSIS

This section focuses on the setup of the sensitivity analysis. It has been designed according to discussions with several Swiss and Costa Rican DSOs. Notably, the study assumes that advanced metering devices are installed at the feeder, all industrial loads, and all PV systems, as it becomes standard in modern distribution networks. However, only a portion of residential and commercial end-users are equipped with a smart meter. The yearly (or monthly) energy consumption of all end-users is still known to the system operator as it is required for billing purposes. Alternatively, their average energy consumption can be estimated based on statistical surveys. Besides, the load flows and voltages might be measured at the MV/LV transformer and cable distribution cabinets. Since PMUs are very rarely installed in LV grids, their measurements are not considered in this chapter. Two strong assumptions are finally made. First, a fully digitized and reliable network model must be available. Although this cannot be taken for granted, it becomes reasonable to consider that the topology and parameters of LV grids can be digitized as explained in Section 2.4.2. In addition, current topology verification and correction algorithms allow for satisfactory grid model quality [109–111]. Second, defective measurement devices and faulty measurement data are assumed to be filtered out beforehand. This necessarily presupposes a comprehensive data preparation as presented in Section 3.3. In any case, preliminary data cleaning cannot systematically eliminate faulty measurements. Advanced error detection and correction steps in the SE process should ideally be taken into account but are out of the scope of this thesis.

In the following, Section 7.3.1 first introduces the benchmark grid model which relies on a large real-world network with the corresponding measurement data. The sensitivity analysis considers four dimensions related to the AMI design (hardware) and two dimensions related to pseudo-measurement synthesis (software), each with multiple options, which is detailed in Section 7.3.2. Next, Section 7.3.2 defines the metrics used for assessing the SE performance of the multiple scenarios with respect to the benchmark model. Different metrics allow for the evaluation of different aspects of the performance. Finally, the outcomes of the sensitivity analysis are illustrated for various quantities and evaluation metrics with the help of box and whisker plots in Section 7.4.

7.3.1 *Benchmark Grid Model*

The grid model used as a benchmark corresponds to the actual distribution grid of an area of the City of Basel, illustrated in Figure 3.4. As a reminder, it mainly covers a residential area and consists of 14 MV/LV transformers, 28 cable distribution cabinets, 971 buses, from which 492 buses are loaded, and 976 lines, from which 12 lines are at the MV level. At the time of data preparation, the grid connected 2610 residential and commercial consumers, 11 industrial customers, and 17 PV systems. It has been chosen for its good data quality and availability, which is not common for such a large real-world grid. Nevertheless, the GIS model of the grid, as well as the corresponding measurements, are single-phase. Although the load in the distribution grid is certainly unbalanced, the state estimation must rely on a single-phase setup. More information on the actual measurements, the characteristics, and the representativeness of the benchmark grid model is provided in the following.

7.3.1.1 *Measurements*

All industrial end-users and PV systems are equipped with a smart metering device providing active and reactive measurements, which is also assumed in the sensitivity analysis. However, only 962 residential and commercial end-users actually had a reliable smart meter at the time of data preparation, which covered about 40% of the total load from this sector. In order to obtain a complete realistic benchmark model, all non-metered consumers have been assigned with active power load profiles of existing smart metered consumers with similar energy consumption from other areas of the City of Basel. Reactive power consumption was unfortunately not recorded by smart meters. Hence, reactive power measurements are synthesized according to the adaptive Power Factor (PF) approach presented in Section 6.2.3.3. Active and reactive power measurements from a confidential data set with similar types of consumers have been provided by Adaptricity (see Section 2.5.1) to serve as training data for the lookup table. For more variability, up to 5% noise is randomly added to the adaptive PF values. It is clear that this process does not lead to reactive power measurements which are faithful to reality. As mentioned in Section 6.3.2, this is not feasible based on the available information. Nevertheless, the process aims to reproduce a realistic distribution of reactive power values. Besides, the temporal resolution of all measurement data is 15 minutes, which is standard in current distribution grids. Note that synchronization delays due to the communication system are not an issue at this temporal resolution.

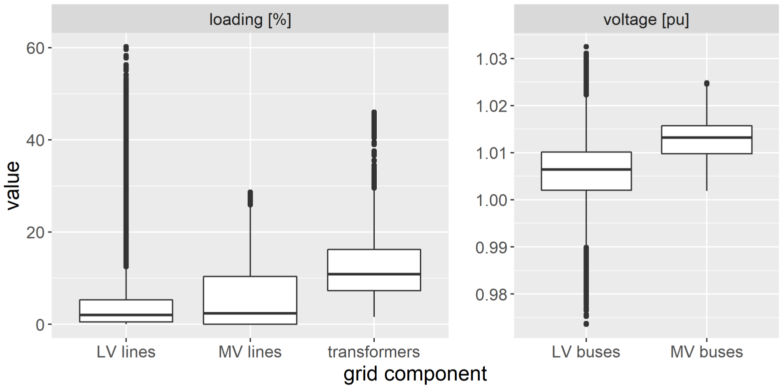


Figure 7.2: Component loading and voltage of the benchmark grid.

The knowledge of the power injection at each bus and of the voltage at the feeder bus provides just enough measurements to obtain an observable system. Hence, power flows and voltages are determined by load flow simulation with the help of Matpower in Matlab [320]. The power losses in the system (e.g., over the lines and transformers) are estimated at 2% of the total load by IWB, the DSO of the City of Basel. Finally, a uniformly distributed noise up to ±0.1% and ±1% is added to the voltage and power quantities, respectively, to mimic the accuracy of the metering devices. Finally, as input for the state estimator, all quantities are converted to the per unit (pu) system, with a base power value of 1 kVA.

7.3.1.2 *Characteristics*

Figure 7.2 illustrates the component loadings and the voltages observed in the benchmark grid after load flow simulation. It appears that the grid is characterized by significant capacity reserves for its current load as the loading of all LV lines, MV lines, and transformers does not exceed 60%, 30%, and 50% of the corresponding capacity, respectively. Similarly, the voltage always lies within the acceptable range of 1 ± 0.05 pu.

Although the grid chosen as benchmark appears as over-dimensioned and can perfectly withstand the levels of component loading and voltage variations with respect to the current load, this is not necessarily valid for all distribution grids in general. In any case, a proper analysis of the penetration and combination of AMI measurements, as well as adequate modeling of the different quantities in distribution grids, is crucial. Notably, the provision of ancillary services such as reactive power and voltage control via demand response or a battery management system requires a good estimate of the state of the grid. As the sensitivity analysis is carried out in relative terms in this chapter, the outcomes are applicable to other urban distribution grids which might not be as strong as the benchmark grid. In that case, the peak load and voltage variations have to be accurately assessed for proper congestion management. Furthermore, an increasing share of Distributed Energy Resources (DERs) such as PV systems, heat pumps, and electric vehicles characterized by high peak consumption is expected to push the infrastructure of current distribution grids to the limits [321, 322]. Cost-effective integration of DERs is only possible with accurate knowledge of the grid's state, and particularly the extreme values.

7.3.2 *Dimensions*

The sensitivity analysis is based on a combination of six dimensions, four of which depend on the metering infrastructure, namely the penetration level, the type of metering devices, their placement, and their capability (i.e., measured quantity). The remaining two dimensions refer to the modeling of pseudo-measurements for active and reactive power injection. Each dimension consists of two to three options. A scenario corresponds to a given combination among the different options. Table 7.2 lists all dimensions with their possible options, and Figure 7.3 presents the structure of the sensitivity analysis, which leads to a total of 144 distinct scenarios. If the AMI hardware setup generates a redundant set of measurements, DSSE can be carried out to estimate the most probable state of the system. In this case, the outcome depends above all on the pseudo-measurement synthesis and on the DSSE algorithm itself. In contrast, there is no redundancy in the absence of both smart meter voltage measurements and measurements at the transformers and cabinets. The sole availability of bus power injections based on SM measurements and pseudo-measurements only allows for a load flow simulation, where the grid state is directly given by the set of independent measurements. Detailed information on each dimension is given in the following.

7.3.2.1 *Penetration of Measurements*

As typically seen in current distribution grids, different penetration levels of smart meters are considered, going from 25% to 75% SM coverage. Full penetration of reliable smart meters leads to an outcome extremely close to the benchmark model. In this case, only the error margin of measurement

Dimension	Options
SM penetration	$25\%, 50\%, 75\%$
Metered grid components	feeder, transformers, cabinets (incremental)
Active power synthesis	AMCM, SLA
Reactive power synthesis	adaptive PF, average PF
SM placement	random placement, strategic placement
SM capability	voltage information, no voltage information

TABLE 7.2: Dimensions of the sensitivity analysis.

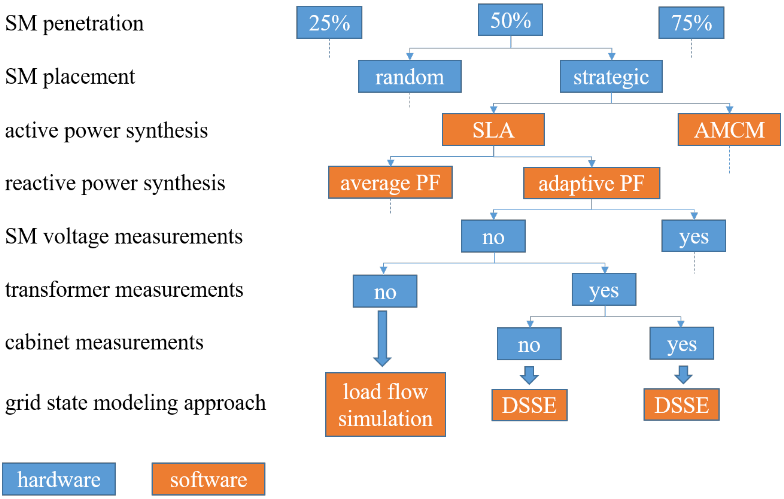


FIGURE 7.3: Structure of the sensitivity analysis according to the different dimensions and options under consideration.

devices can induce small deviations which are negligible in comparison with a partial SM coverage where the SE errors are principally induced by pseudo-measurements. Hence, full SM penetration is not explicitly analyzed in this chapter. Besides, note that faulty measurements can be seen, to some extent, as pseudo-measurements with a decrease of the actual SM penetration. As it can be observed in Figure 7.4 at a representative bus, a higher SM penetration obviously leads to more accurate state estimation. More than the sole impact

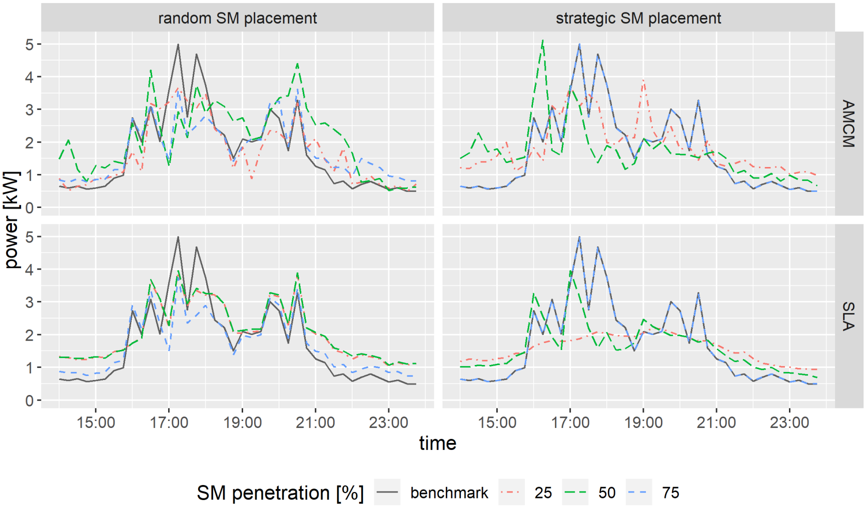


Figure 7.4: Example of active power injection at a particular bus with respect to different SM placements, SM penetrations and types of pseudo-measurements.

of the SM penetration level, its combination with other dimensions is of interest in this sensitivity analysis.

In addition to SM measurements, the AMI can also include measurements taken at the feeder, at transformers, and at cable distribution cabinets. They are labeled as grid component measurements. This study assumes that the feeder is metered in any case and that cable distribution cabinets can be equipped with a metering device only if transformers are also metered. Hence, three options are examined, where transformer and cabinet measurements are incrementally added to feeder measurements. For all metered grid components, the voltage and all associated power flows are assumed to be recorded.

7.3.2.2 *Synthesis of Pseudo-Measurements*

As previously mentioned, pseudo-measurements are indispensable in DSSE to cope with the lack of direct measurements. Presented in Section 6.2.2, the Standard Load Allocation (SLA) and the Adaptive Markov Chain Model (AMCM) with optimal aggregate matching are the two approaches considered for active power pseudo-measurement synthesis. As a reminder, the AMCM is a stochastic load profile generator, and the optimal aggregate matching

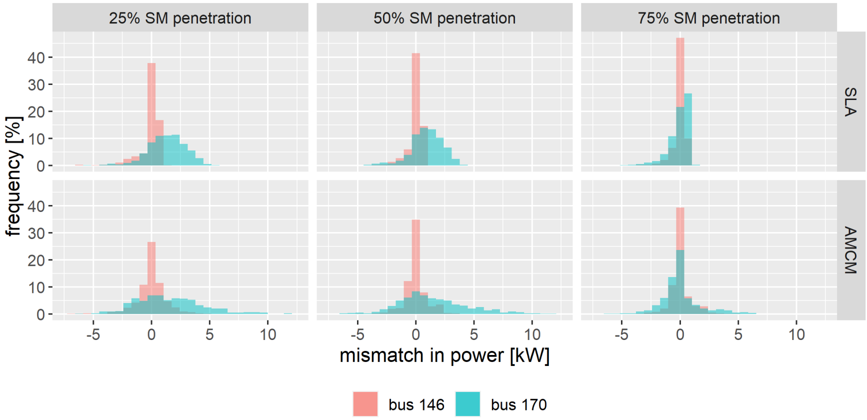


FIGURE 7.5: Error distribution of active power injection (before SE) at two specific buses between multiple scenarios with random SM placement and the benchmark model, considering different SM penetration levels and two approaches for active power pseudo-measurement synthesis.

consists of a load profile selection and a load profile allocation step. These are under-constrained problems, which results in multiple feasible solutions, and consequently, different possible power flows for a single time step. However, only one realization per non-metered consumer is taken into account. In the sensitivity analysis, the idea is to compare an approach traditionally used by DSOs with an innovative method that faithfully represents the load distribution visible in SM data. Figure 7.4 shows that spiky active power profiles are created by the AMCM-based approach, whereas smoother profiles result from the SLA. Furthermore, Figure 7.5 illustrates the error distribution of active power injection at two different buses before SE. The general shape of the distribution highly varies between the two examples, but also according to the type of active power synthesis. For example, the error distribution at bus 146 for a 25% SM penetration is left-skewed with SLA-based synthetic load profiles but exhibits a Gaussian behavior with SLA-based synthetic load profiles. The SM penetration level mainly impacts the variance of the error distribution. Note that the large majority of pseudo-measurements proposed in the DSSE literature have similar statistical properties as SLA-based load profiles.

Regarding reactive power pseudo-measurements, the average and adaptive PF approaches as defined in Section 6.2.3 are considered. It must be reminded that the adaptive PF approach has also been leveraged to create reactive power injections for the benchmark model due to the absence of direct reactive power measurements recorded by smart meters. Although the input data for creating the respective lookup tables are different, a direct comparison with the average PF approach might be biased. Hence, the purpose of this analysis is not to identify which is the most suitable approach but to study the influence of reactive power synthesis on the DSSE outcome. In addition, the performance evaluation will not directly focus on reactive power quantities.

7.3.2.3 *SM Placement and Capability*

With a limited number of smart meters, DSOs face the challenge of their placement across the grid. The basic approach consists of randomly distributing the smart meters. Alternatively, they can be wisely allocated to end-users in order to maximize their benefits. The theoretical background of optimal meter placement is largely investigated in the literature and is summarized in [114]. The optimal solution inevitably depends on a certain objective, e.g., improving system observability [323], minimizing installation and maintenance costs [324, 325], or improving the DSSE accuracy [314, 316, 326]. In addition, the optimization problem is generally solved by heuristic search, Genetic-Algorithm (GA), or Mixed Integer Linear Programming (MILP), which can be computationally intensive on large real-world grids. In this sensitivity analysis, the random SM placement is compared with the so-called strategic placement approach. This is a simple heuristic approach where smart meters are installed at the end-users with the highest energy consumption. The strategic approach might be sub-optimal depending on the objective but still provides very satisfying results and can be easily implemented knowing the average energy consumption of each end-user.

Besides, only one representative scenario of random placement is considered. This is justified by Figure 7.6 which compares the median error over all nodal active power injections based on ten different random SM placements per scenario. The error is expressed as adjusted RMSE which is defined in Section 7.3.3.2. It appears that the variations among the different random placements are minor, especially with respect to the strategic placement, the penetration level, and the type of pseudo-measurement synthesis. As an illustration, Figure 7.4 shows that the strategic placement allows for a better estimation of the total load at this specific bus since smart meters cover a higher share of power consumption than with random SM placement.

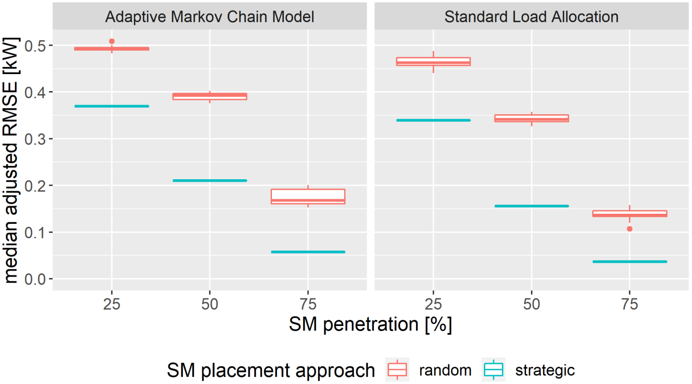


FIGURE 7.6: Comparison of the median adjusted RMSE in active power injections between the strategic and 10 random SM placements for different levels of SM penetration, and different types of active power pseudo-measurement synthesis.

Notably, a 75% penetration of strategically placed smart meters results in a full load coverage at this specific bus (i.e., all consumers at this bus belong to the 75% of end-users with the highest energy consumption and are assigned with a smart meter).

Furthermore, the choice of the SM placement approach influences the parametrization of Equation (6.14) when it comes to matching optimally individual AMCM-based synthetic load profiles with aggregate power measurements. With a random placement of smart meters, synthetic load profiles tend to have similar statistical properties (e.g., mean energy consumption) as the non-metered consumers. Hence, the scaling factor α in Constraint 6.14b is set to 1 in order to guarantee the allocation of at least one synthetic profile per non-metered consumer on average. However, the strategic SM placement induces the synthesis of load profiles that all have a higher mean energy consumption than the non-metered consumers[3]. Accordingly, α is set to a value of 0.8 to prevent an overestimation of the total energy consumption of non-metered consumers. In this case, this implies that approximately 20% of the non-metered consumers (i.e., the consumers with the lowest energy consumption) are assigned with a zero load profile.

---

[3] It must be reminded that synthetic load profiles are generated based on the available smart meter data in the same system.

The last dimension considered in the sensitivity analysis refers to the voltage measurements of smart meters. Although all smart meters have to measure the voltage to obtain power quantities, that information is not automatically recorded and transmitted. From the perspective of power companies, active power is usually the main quantity of interest, and further data require additional communication and storage capacity. Therefore, the added value of the knowledge of voltage measurements is also studied.

7.3.3 *Evaluation Metrics*

Special attention must be given to the choice of the evaluation metrics. In fact, each metric looks at the performance of a given approach from a different perspective, which provides additional information. Different metrics are usually complementary. As studied in Chapter 5, the load of a single consumer or a small aggregation of consumers in distribution grids is particularly volatile, meaning drastic changes from one time step to the next. Spikes in active and reactive power consumption can be several dozen times higher than the mean power consumption. Largely used in literature for the evaluation of DSSE outcomes, the common Root-Mean-Square Error (RMSE) seems appropriate for relatively stable signals but gives a biased and inadequate measure of the estimation accuracy of volatile profiles. In this work, the point-wise accuracy of power injection and power flow estimates is evaluated via an adjusted version of the RMSE that does not directly penalize large point-wise errors. Furthermore, the ability to properly represent extreme values is assessed based on the notion of 95^{th} percentile. Since the voltage fluctuates around 1 p.u. and is not highly volatile, it is still evaluated via the common RMSE.

7.3.3.1 *Common Root-Mean-Square Error*

The common RMSE is a widely used metric that considers the point-wise error between two time series:

$$\text{RMSE} = \|y - \hat{y}\|_2 = \sqrt{\frac{1}{T} \cdot \sum_{t=1}^{T} (y_t - \hat{y}_t)^2}, \tag{7.3}$$

where $y = (y_1, \ldots, y_t, \ldots, y_T)$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_t, \ldots, \hat{y}_T)$ are the true and estimated time series, respectively, and T is the number of time steps under consideration. The lower the RMSE, the better the estimation. The RMSE has the same unit as the evaluated quantity. Due to its square function, the common RMSE penalizes more large point-wise errors.

7.3.3.2 *Adjusted Root-Mean-Square Error*

As detailed in Section 9.3, the common RMSE leads to the so-called double penalty effect when comparing two volatile time series. The double penalty effect occurs in the estimation of spikes, where the magnitude is properly predicted, but its point in time is slightly displaced. In this case, the common RMSE penalizes both the time step when the true spike is not estimated and the neighboring time step associated with the wrongly estimated spike. This phenomenon is illustrated in the top-right subplot of Figure 7.4 with a level of 50% SM penetration. As mitigation measure, the authors in [136] propose an adjusted error metric that allows for small, possibly discontinuous, displacements of the estimated values in time. The adjusted error is the solution of the following optimization problem:

$$\begin{aligned} \mathrm{RMSE}^{\omega} = \min P \in \mathcal{P} \quad & \|y - P \cdot \hat{y}\|_2, && (7.4\text{a}) \\ \text{s.t.} \quad & P_{uv} = 0, \quad \forall\, |u - v| > \omega, && (7.4\text{b}) \end{aligned}$$

where $\omega \geqslant 0$ is an adjustment limit, P is a permutation matrix, $P_{uv} \in P$ refers to the displacement of the estimated value $\hat{y}_u$ from time step u to time step v, and $\mathcal{P}$ is the complete set of restricted permutations. The optimization function (7.4a) is similar to Equation (7.3) for the common RMSE, except that the estimated values can be permuted across time steps in order to minimize the resulting error. The error minimisation is a variant of the assignment problem, a well-known combinatorial optimisation problem which can be solved using the so-called Hungarian method. More details are provided in [327]. The equality constraint (7.4b) ensures that each estimated value is not displaced more than the adjustment limit ω. In fact, the adjustment limit is incorporated into the algorithm by setting:

$$|y_u - \hat{y}_v|^2 = C, \quad \text{if } |u - v| > \omega, \tag{7.5}$$

where C is a large constant that prevents such displacement. As discussed in [136], the value of the adjustment limit highly impacts the outcome of the metric. If $\omega = 0$, the common RMSE is recovered. An increase of the value of ω reduces the adjusted error. Nevertheless, a small error resulting from large displacements is not necessarily indicative of a good estimation such that a compromise has to be found. For this sensitivity analysis, ω is set to 4 hours.

7.3.3.3 *Ninety-Fifth Percentile*

The 95th percentile is a threshold value below which 95% of the observations fall. On this basis, following error metric is proposed to assess whether the peak values of a time series are well estimated:

$$e^{95} = \hat{y}^{95} - y^{95} \tag{7.6}$$

where e^{95} is the proposed error metric, and $\hat{y}^{95}$ and $\hat{y}^{95}$ are the 95th percentile values of the estimated and true time series, respectively. In this case, the closer to zero, the better the estimation. A metric value lower or higher than zero reflects an underestimation or overestimation of the peak values, respectively.

7.4 RESULTS OF SENSITIVITY ANALYSIS

The sensitivities of active power injections, line loadings, and voltages with respect to the six dimensions given in Section 7.3.2 are discussed in this section. Although only the voltage and sometimes active power injections are traditionally considered in DSSE literature, the loading of power lines and transformers is also of high interest to DSOs, especially regarding the risk of overloading. For each scenario, a time series simulation over one week is carried out. Based on 15-minute resolution data, the resulting 672 time steps cover a wide range of realistic cases and serve as a practical alternative to Monte-Carlo simulations. The outcomes are illustrated in the form of box and whisker plots. Each figure of this section displays the performance evaluation of all 144 scenarios according to a specific metric and a specific quantity. In order to represent all six dimensions in the same figure, the subplots are organized as follows:

- the rows refer to the type of active and reactive pseudo-measurement synthesis,
- the columns refer to the SM placement and capabilities,
- the x-axis refers to the level of SM penetration,
- the colors refer to the measurement of grid components.

This allows for a comprehensive overview of all possible combinations. In this way, the reader can navigate vertically and/or horizontally within the figures and directly see the impact of switching options in the pseudo-measurement synthesis and/or the AMI design, respectively. Each data point represents the corresponding metric value for a single bus or line and for one scenario over the entire considered time frame. The central bar in a box and whisker plot indicates the median value, the box corresponds to the Interquartile Range (IQR), and the ends of the whiskers define $1.5 \cdot \mathrm{IQR}$ below and above the lower and upper quartiles, respectively. For visualization purposes, all outliers (i.e., values beyond the whisker ends) are discarded.

7.4.1 *Power Injection*

Based on the adjusted RMSE, the share of smart meters and their placement across the grid are the main dimensions that play a role in the estimation of active power injections, as illustrated in Figure 7.7. For example, a strategic SM placement allows for a similar accuracy in average as a 25% higher share of smart meters being randomly allocated. Additional grid component measurements, as well as the voltage information of smart meters, are not particularly beneficial in this case. In other words, the point-wise estimation of active power injections largely depends on the direct measurement of these active power injections and is barely influenced by the measurements of other quantities such as active power flows at a spatial aggregate level. Furthermore, the synthesis of active power pseudo-measurements via the AMCM approach seems slightly worse in terms of point-wise accuracy. In fact, single (or a small aggregation of) residential load profiles are known to be extremely hard to estimate at each time step properly. Hence, the AMCM still produces peak values in time periods where the actual load is generally low. This is highly penalized by the adjusted RMSE, which is still a point-wise metric.

However, by considering the 95th percentile, Figure 7.8 confirms that the SLA generally underestimates the consumption spikes and leads to a higher variance in error than the AMCM-based load profiles. For example, active power spikes suffer a median drop of 29% with a 25% penetration of randomly allocated SMs without cabinet measurements, which is substantial. It must also be noted that the analysis is performed with 15-minute resolution data. As studied in Chapter 5, the drop is expected to be even more significant at higher temporal resolutions. In comparison, the median drop is only 0.8% when active power pseudo-measurements are synthesized by the AMCM approach. Besides, a strategic allocation and a higher penetration of smart meters considerably improve the estimation of peak power values. Especially, they are almost perfectly estimated with a 75% SM penetration, independently of the pseudo-measurement synthesis approach. A higher penetration level

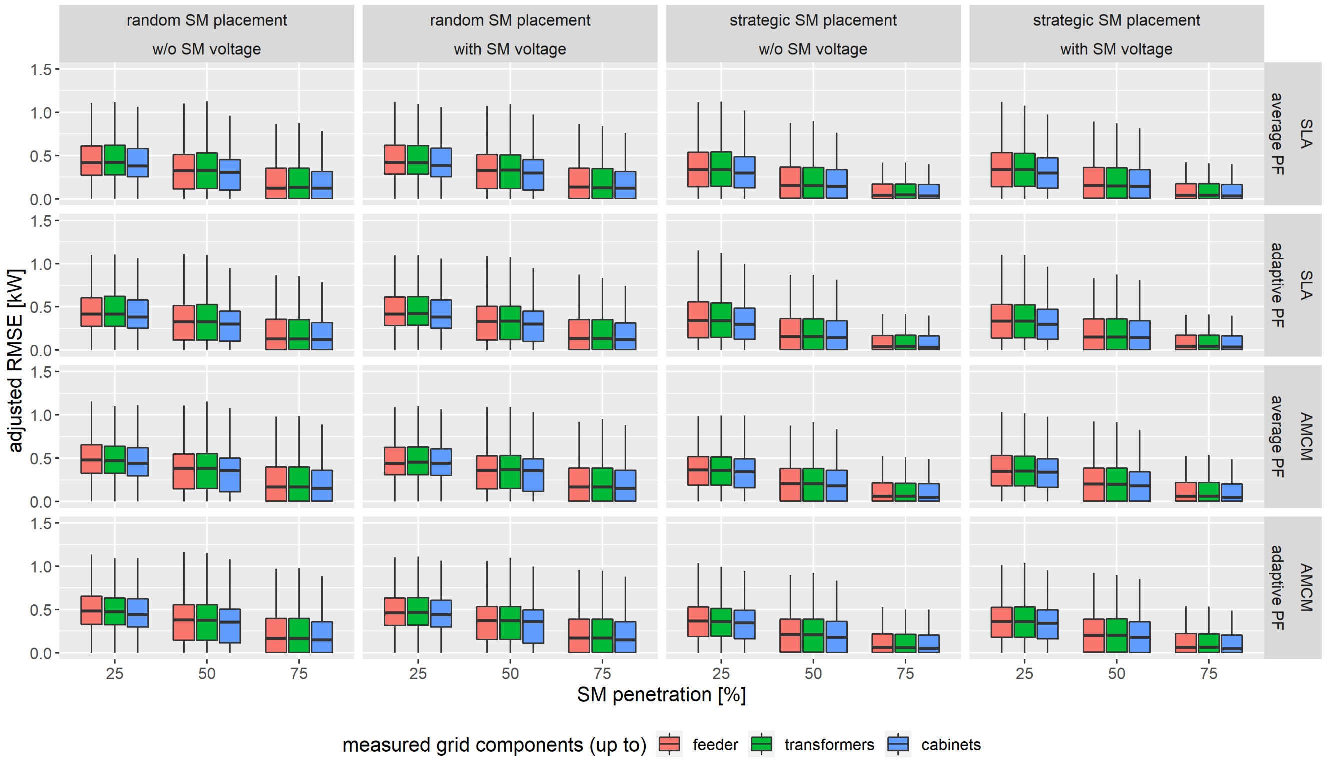


FIGURE 7.7: Adjusted RMSE of active power injections among the loaded buses with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements.

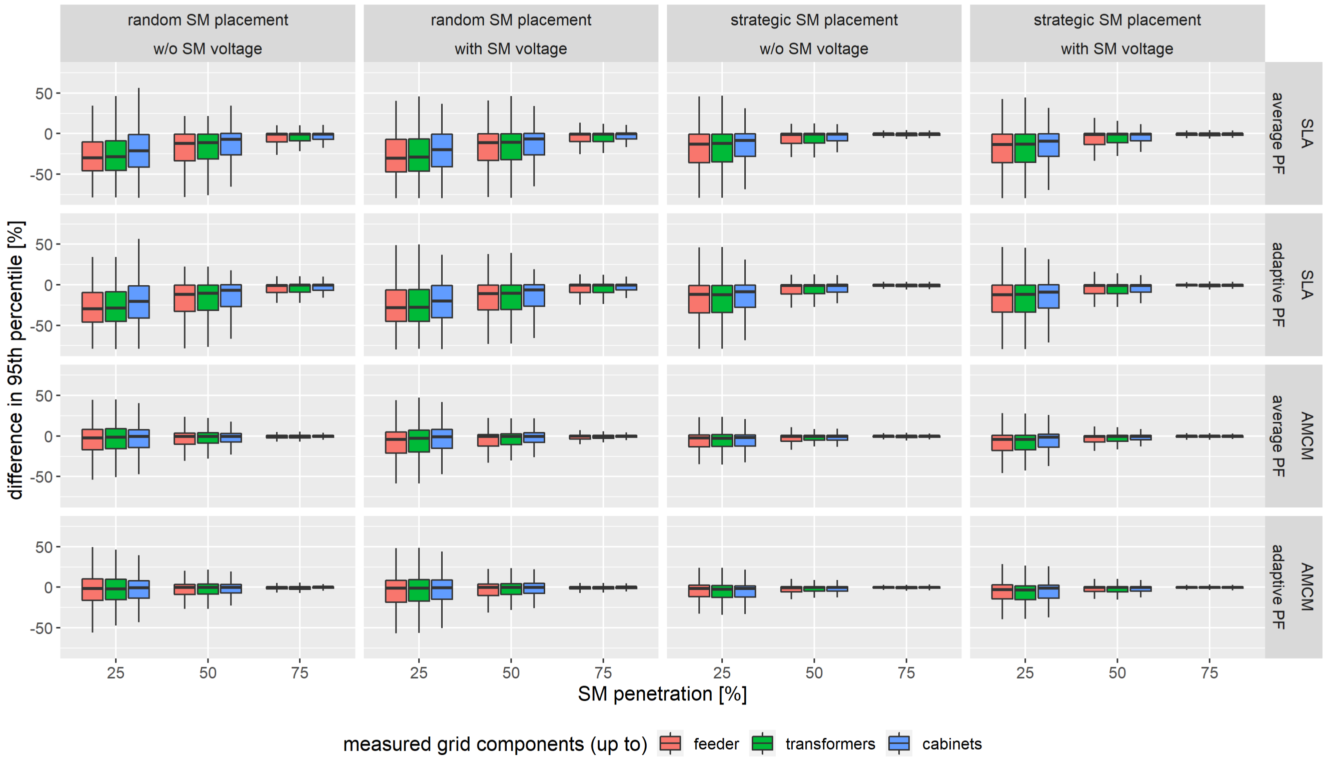


FIGURE 7.8: Difference in 95^{th} percentile of active power injections among the loaded buses with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements.

is not required for a better estimation of peak power values. Finally, the synthesis of reactive power pseudo-measurements has essentially no impact on the estimation of active power injections, both in terms of adjusted RMSE and 95^{th} percentile.

7.4.2 *Component Loading*

The loading of components is expressed in percentage with respect to their maximum capacity. Due to their high capacity, the errors in loading estimation are relatively low (i.e., mostly below 1%). Nevertheless, the sensitivity analysis reveals substantial differences between the scenarios. On the one hand, transformer loading estimation is highly enhanced by the sole addition of a metering device at the transformers. This is obvious and not explicitly illustrated in this chapter. On the other hand, the loading estimation of power distribution lines depends on multiple dimensions. As shown in Figure 7.9, the share of smart meters and their placement approach have a similar influence as in the estimation of active power injections. This is comprehensible since the loading of lines is mainly impacted by active power injections. Moreover, direct measurements of power flows at the cable distribution cabinets improve line loading estimation in general. This is not the case for direct measurements at the level of transformers. In addition, the voltage information of smart meters reflects the aggregate behavior of end-users in a certain neighborhood and is also beneficial for accurate line loading estimation. Besides, the synthesis approach for reactive power pseudo-measurements has a negligible impact in comparison with other dimensions. Although not explicitly illustrated in this section, the 95^{th} percentile metric leads to similar results as in Figure 7.8, but to a lesser extent. Notably, the SLA causes a median drop of 10.3% in the magnitude of spikes in line loading with a 25% penetration of randomly allocated smart meters without cabinet measurements. The misestimation gradually vanishes with increasing SM penetration level.

7.4.3 *Voltage*

The accuracy of the voltage estimates is displayed in Figure 7.10 by means of the common RMSE. The state estimation errors generally lie below 0.2%, which is particularly low. This can be partially explained by the robustness of the distribution grid under consideration and the absence of faulty measurements. The voltage information of smart meters obviously leads to much smaller errors. If only a few voltage measurements are available (e.g., at the

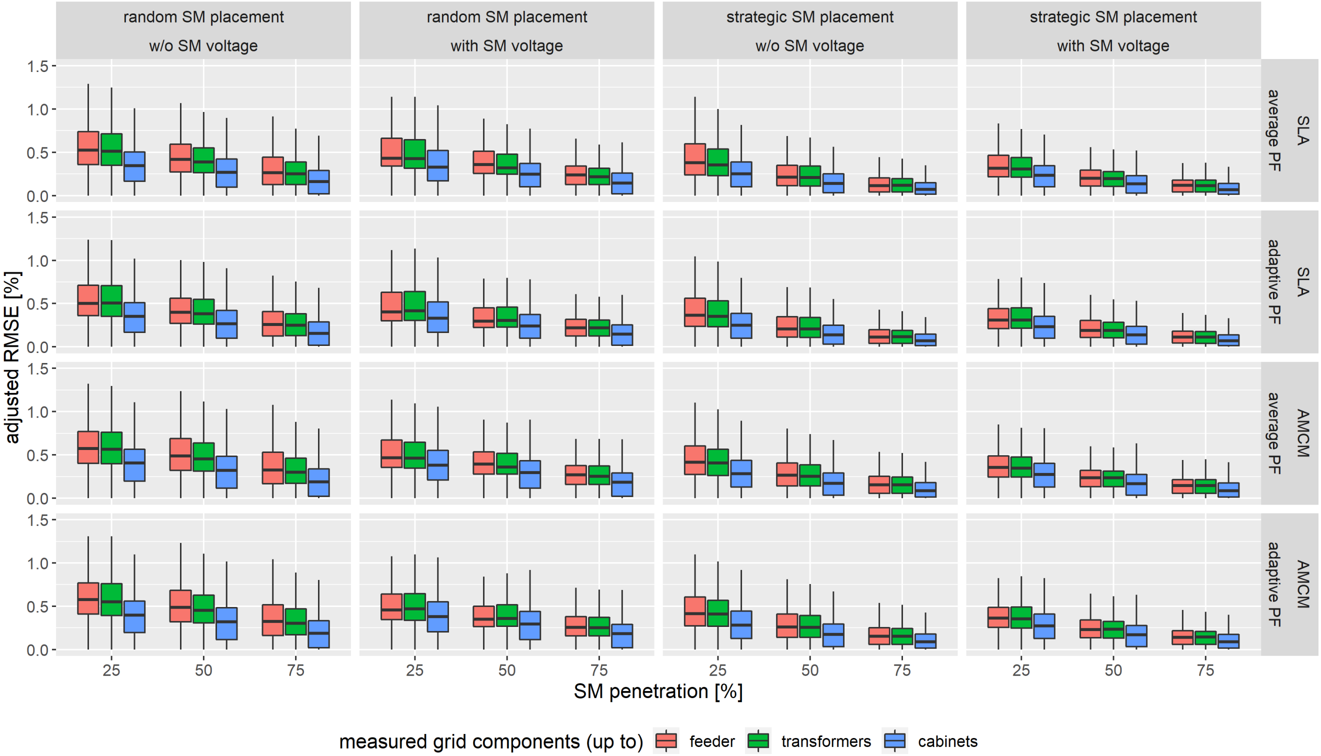


FIGURE 7.9: Adjusted RMSE of line loadings with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements.

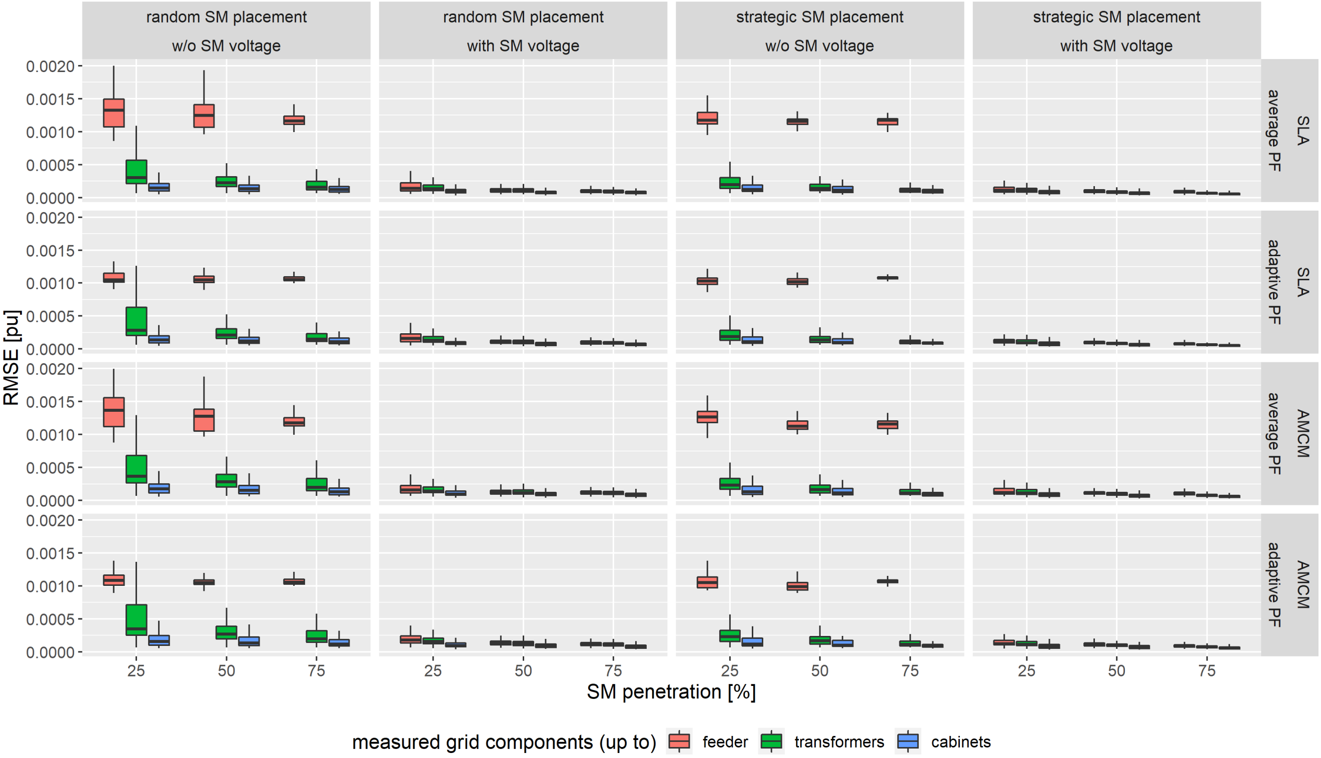


FIGURE 7.10: RMSE of bus voltages with respect to different types of pseudo-measurement synthesis, SM placements and capabilities, and penetrations of measurements.

feeder and at the transformers), the estimation is particularly sensitive to the corresponding noise. In any case, voltage measurements at the distribution cabinets result in very good accuracy for all bus voltages. They enable to do without SM voltage measurements at the end-user level since the voltage is locally very similar. Besides, the strategic SM allocation further improves the voltage estimation compared to the random allocation when cabinet or SM voltage measurements are not available. Finally, the synthesis of reactive power does not play a role in the voltage estimation.

7.5 COMPUTATIONAL COMPLEXITY AND SCALABILITY

In this section, the computational complexity, scalability, and generalization of the sensitivity analysis to other distribution systems are discussed. First of all, it has to be emphasized that the use of real distribution grids, including the LV level with the corresponding SM data, is extremely rare in the academic literature in general. The pieces of work in [295, 328, 329] are among the very few examples which rely on real systems for DSSE-related studies at the LV level. Nevertheless, the grid used in their case study is not larger than a few dozen buses, and the smart meter data do not always come from the same system. The study presented in this chapter demonstrates the feasibility of a comprehensive sensitivity analysis related to DSSE on a large real-world distribution system with almost 1’000 buses and more than 2’600 end-users.

Although the sensitivity analysis is an offline process whose computational time and resources are not as critical as for operational purposes, it must still be computationally tractable. In this study, all simulations are carried out on a 64-bit Windows server with an Intel Xeon Gold 6154 CPU at 3 GHz and span over a period of one week with 15-minute resolution data (i.e., 672 time steps). The synthesis of pseudo-measurements requires less than 20 MB of data from the benchmark system to produce 700 MB of load profiles in about 5 hours for all scenarios. A large majority of the resources are used by the AMCM approach. Furthermore, the state estimation itself lasts about one minute per time step. It results in 90 MB of data based on less than 10 MB of input data per scenario. Obviously, the computational complexity largely depends on the number of scenarios, the simulation period, and the time resolution, which can be adapted according to the DSO’s needs. This chapter precisely considers a large variety of scenarios from which only a fraction might be of interest to a specific DSO, depending on its AMI development strategy. As detailed in Section 3.3, it must also be reminded that the creation and validation of the grid model as well as the gathering, preparation, and cleaning of the measurement data and metadata is an essential prerequisite that is considerably more time-consuming and resource-intensive than the analysis itself.

Based on the above statements regarding the computational complexity, a similar sensitivity analysis is perfectly scalable to larger systems (e.g., at a city level), assuming that the grid model and measurements are already available in a clean and tidy form. The feasibility of state estimation techniques on very large synthetic distribution grids has been proven by the authors in [330–332]. The availability of real grid models, including the LV level, and the preparation and integration of measurement data are the actual bottlenecks for proper transparency down to the end-users in distribution grids. Besides, the presented case study relies on a network that principally supplies residential end-users. It is sufficiently large to extrapolate the outcomes to other similar systems. Although the main tendencies of the analysis can be generalized, the magnitude of the sensitivity results regarding power flows and voltages still depends on the structure and robustness of the grid under consideration as well as the type of end-users. Such a comprehensive sensitivity analysis at a city level is not necessary, but various characteristic sectors (e.g., residential, commercial, and industrial areas) should still be considered to give a representative overview of the entire distribution grid.

7.6 CONCLUSION

This chapter aims to provide realistic insights into the most relevant dimensions to consider for cost-effective data-based modeling of distribution grids, including the LV level. For that purpose, the impact of different combinations and penetrations of state-of-the-art AMI sensors on the WLS state estimation of a large real-world distribution grid is studied in a comprehensive sensitivity analysis. The presented setup accounts for data streams that DSOs can realistically obtain from their AMI. In the case of partial penetration of smart meters at a certain bus, an approach is proposed to weigh the different power measurements in the WLS formulation. Furthermore, novel methods for synthesizing active and reactive power are compared to standard approaches widely used by DSOs. A simple but effective strategic placement of smart meters to the end-users with the highest energy consumption is also compared to random SM placement. Moreover, specific grid component measurements (i.e., at the transformers and cabinets) are taken into consideration in addition to SM data. Among others, the performance evaluation relies on an adjusted error metric that does not unfairly penalize volatile quantities. The representation of peak load is further assessed based on the 95th percentile.

In general, the efficiency of a certain AMI configuration and of a certain approach for pseudo-measurement synthesis depends on the evaluation metric and on the quantity under consideration. Obviously, a higher SM penetration and their strategic placement substantially increase the SE accuracy in any case. Notably, an SM penetration level higher than 75% is not justified, especially if smart meters are strategically placed. It also appears that additional measurements at the transformers and cable distribution cabinets barely influence the estimation of active power injections. They are nevertheless helpful when it comes to estimating the voltage and line loading. Also, voltage measurements given by smart meters are not particularly beneficial if distribution cabinets can provide this information. Moreover, smooth synthetic pseudo-measurements, as generally considered by DSOs and in the existing DSSE literature, clearly underestimate the magnitude of spikes in power injection and in line loading in case of a low SM coverage. This can be resolved by the use of realistic load profiles, e.g., generated by the proposed AMCM approach. They inevitably lead to a slight decrease in point-wise accuracy but realistically reflect the load distribution in the system. Besides, a substantial influence of the synthesis approach for reactive power pseudo-measurements cannot be observed in the proposed setup. However, the topic of realistic synthesis of reactive power profiles merits closer scrutiny.

To sum up, it is particularly relevant to model distribution grids down to the LV level based on representative data. In fact, some DSOs currently experience unexpected contingencies (e.g., congestions at peak hours, voltage band violations) due to the misestimation of non-observable quantities in their LV grid. Among others, this highlights the importance of proper pseudo-measurement synthesis. Moreover, future works must definitely consider three-phase measurements and network models since distribution grids are typically unbalanced. It is highly probable that the level of volatility observed in the single-phase system is somewhat reduced by the aggregation effect over the three phases. In the same vein, the use of measurement data at higher temporal resolution (e.g., one-minute resolution) better reflects the actual volatility. Additional quantities such as PMU measurements might also become standard in future distribution grids, which potentially contributes to increasing the system observability. In any case, the use of appropriate evaluation metrics is an essential part of the solution, which should not focus purely on point-wise accuracy. These different aspects must be kept in mind and further analyzed with the increasing penetration of DERs and electric vehicles whose stochastic nature adds even more uncertainty.

Furthermore, the WLS algorithm is probably not the most suitable approach considering the characteristics of state estimation in distribution grids. Especially, the WLS algorithm assumes a Gaussian error distribution, which is generally not valid in LV grids as illustrated in Figure 7.5. In addition, the authors in [333] highlight that uncertainties in both the parameters of the grid model and in the measurement data influence the performance of DSSE algorithms. Hence, alternative DSSE approaches specifically designed to cope with a large share of leverage points (i.e., measurements that significantly influence the SE solution), pseudo-measurements, and possible faulty measurements, and that can take grid model uncertainties into consideration are more appropriate than the WLS algorithm. In this context, Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF), Least Absolute Value (LAV), Forecasting-Aided State Estimation (FASE), and various ML techniques are reviewed by the authors in [8, 114].

Eventually, probabilistic concepts should be preferred to deterministic approaches in order to account for the large uncertainties induced by the lack of direct measurements. For example, the authors in [329] propose a promising probabilistic low-voltage state estimation using analog-search techniques based on the historical data. It must be noticed that the error in DSSE mainly originates from pseudo-measurements. Nevertheless, statistically representative pseudo-measurements provide insight into the actual uncertainty. Notably, the proposed AMCM is a stochastic algorithm that inherently accounts for the value distribution and the seasonality observed in smart meter data. Hence, multiple AMCM realizations could be integrated into a probabilistic DSSE framework to determine the state estimation uncertainty. The principle could be similar to the estimation of the uncertainty in wind power forecasting as detailed in [334]. Alternatively, the idea of probabilistic state forecasting is presented in Section 9.4.

Part III

APPLICATION OF DATA-BASED APPROACHES IN LOW-VOLTAGE GRIDS

8

DETECTION AND DISAGGREGATION OF FLEXIBLE RESIDENTIAL LOADS

This chapter proposes novel unsupervised approaches for the detection and disaggregation of cold appliance and water heater loads from standard smart meter data. In a context where increased flexibility is required from the demand side, a non-negligible potential is seen in the residential sector. For example, cold appliances such as refrigerators are continuously active and potentially available for demand response. Besides, electric storage water heaters are appliances with relatively high power consumption and significant thermal inertia. Nevertheless, the success of efficient demand response schemes via direct load control presupposes an accurate estimate of their power demand at each instant, not only at an aggregate level but precisely at the device level. Although the load of single devices is rarely directly measured, a large penetration of smart meters enables to indirectly infer this information via load detection and disaggregation. In contrast to non-intrusive load monitoring techniques generally suggested in the literature, the proposed detection and disaggregation approaches are unsupervised and only rely on commonly available smart meter measurement data with a temporal resolution between 1 and 30 minutes. Their applicability is demonstrated in 70 households. This chapter is based on [335] and [336] for the disaggregation of cold appliances and water heaters, respectively.

8.1 INTRODUCTION

In future distribution grids, Demand Response (DR) is seen as an efficient way to integrate an increasing share of intermittent Renewable Energy Sources (RES) as the low controllability of RES can be compensated by the flexibility potential on the demand side. In a DR scheme, some flexible loads are required to shift or decrease their consumption. This potential can be exploited via price signals which discourage consumption at high loading conditions and possibly encourage consumption in periods with considerable renewable energy production. The industrial sector already profits from such time-varying electricity prices. Alternatively, some electric devices could potentially be directly controlled by system operators without impacting the comfort of the user, which is known as Direct Load Control (DLC). The thermal inertia of Thermostatically Controlled Load (TCLs) such as air conditioners, space heaters, heat pumps, electric water heaters, and refrigerators offers great flexibility, which is particularly interesting for DR applications [159]. In addition, the direct control of TCLs does not require any intervention by the user, at the difference of domestic appliances like washing machines and dishwashers, or even electric vehicles [50, 139]. Still largely under-exploited, the flexibility potential in the commercial and residential sectors is non-negligible. The authors in [158] estimate the hourly average load reduction potential in Europe through load shedding and shifting between 20 GW and 75 GW for both residential and tertiary sectors, depending on the period of the day and of the year. In comparison, the industrial load reduction potential is estimated at 25 GW over the whole year.

In practice, when direct load control is required by the system operator, DLC signals are traditionally blindly sent to all participating end-consumers without knowing their actual flexibility potential. In fact, the flexible devices react and temporarily reduce their consumption only at the condition that this does not affect the end-user's comfort. In order to design more effective DR schemes, the consumption pattern and the availability of the individual devices under consideration must be accurately estimated. For that purpose, multiple approaches are proposed in the literature. Some techniques rely on typical load profiles, associated with statistical surveys about their consumption share in a certain population [337, 338]. Nevertheless, these techniques only give a rough and general estimate of their consumption at an aggregate level and cannot correctly account for the high diversity of individual electrical devices. Besides, a large variety of model-based approaches are developed to approximate the internal physical behavior of individual flexible devices, notably focusing on TCLs [339–341]. Concretely, the electricity consumption of TCLs is either ON or OFF, following a so-called bang-bang controller which aims to keep the associated temperature within limits. Model-based approaches are particularly convenient for designing control algorithms but can hardly reflect the actual behavior of flexible devices in a given real system. They require different modeling parameters (e.g., temperature limits, device specifications, type of perturbations) which are rarely known in reality.

Recently enabled by the wide-scale roll-out of advanced metering devices, data-based approaches provide the means for accurate quantification of the share of flexible loads in a given distribution system down to the end-user level. For example, Intrusive Load Monitoring (ILM) directly provides sub-metering data of the devices of interest, which gives perfect knowledge of their consumption pattern. For example, the Belgian pilot project LINEAR could estimate the flexibility potential of multiple residential appliances based on different DR experiments and the exploitation of sub-metering data [139]. Furthermore, the authors in [342] have leveraged sub-metering data with a 1-minute resolution to decompose the load at an aggregate level (i.e., 50 to 1000 houses) into seven categories of appliances. They demonstrate that the disaggregation accuracy via an ANN is satisfactory with only 5% of the consumers providing their sub-metering data, as long as the aggregation level is reasonably high (i.e., at least 200 users) and the end-users are relatively homogeneous [343]. Nevertheless, measurements of individual appliances are currently only feasible in small-scale experiments. On the one hand, their high installation cost inhibits the generalization of sub-metering devices in distribution grids. On the other hand, ILM is controversial in terms of privacy [344]. In contrast, smart metering at the end-consumer level becomes the norm in an increasing number of distribution systems. Hence, Non-Intrusive Load Monitoring (NILM) techniques are often mentioned in the literature to extract the load profile of single appliances from power measurements at the end-consumer level [345, 346]. They rely on the detection and learning of the typical electric device's signature, i.e., its unique steady- or transient-state characteristics. Often based on HMMs or ANNs, they can be supervised (i.e., they require sub-metering data as training data) or unsupervised [347].

However, commonly proposed NILM approaches in the literature presuppose a sampling rate of at least 1 Hz (or much higher for the detection of transient-state characteristics). This is far from the standards of smart meters, which record measurements at most with a 1-minute granularity, which is generally insufficient to detect the device signature. The literature on load disaggregation based on smart meter data with a sampling period of at least one minute is extremely scarce. For example, the authors in [348] compare the performance of four classifiers (i.e., KNN, random forest, linear SVM, and non-linear SVM) on 1-second and 1-minute measurement data for the event detection of 9 types of devices, including the refrigerator. Alternatively, the authors in [349] compare the event detection performance of a Factorial Hidden Markov Model (FHMM) and a sparse matrix processing based technique on two open data sets over a range of sampling periods from 6 seconds to 15 minutes. The algorithms are used as multi-class classifiers for the event detection of 5 types of devices, also including the refrigerator. Nevertheless, the disaggregation processes proposed in [348] and [349] require sub-metering data in the training phase and do not focus on the load detection but on the event detection (i.e., either ON or OFF). In any case, there is a clear lack of disaggregation approaches based on standard smart meter data[1] in the literature.

This chapter focuses on the detection and disaggregation at the house level of cold appliances, especially refrigerators, and electric storage water heaters based on standard SM data without the need for sub-metering data. First, cold appliances are often disregarded from the current load disaggregation literature because their rated power can be up to 50 times lower than the power of other flexible loads such as the AC unit, the water heater, or the cloth dryer. These other loads would potentially have a higher impact in a DR scheme [148, 287]. Nevertheless, they have the advantage of being present in practically all houses. In the European residential sector, the authors in [158] estimate that almost half of the load reduction can be realized by shifting consumption of freezers and refrigerators. Moreover, the authors in [350–352] demonstrate that effective control algorithms can theoretically be applied to cold appliances. In fact, smart controllable refrigerators are already commercially available [353, 354]. Second, electric storage water heaters are among the appliances with the greatest flexibility potential and thermal capacity [139, 287]. They are commonly the largest consumers in a building in terms of rated power, and their ramp rate from zero to full power is almost instantaneous. In addition, their energy storage potential is available over several hours and can be exploited at any period of the year, as their usage is only marginally impacted by seasonal variations [341].

The remainder of this chapter is structured as follows. Section 8.2 details the methodology behind the detection and disaggregation processes of both the cold appliances and the electric storage water heaters. Although the small steady-state power variations that build the load signature are smoothed out and not detectable at standard smart meter resolutions, some characteristics unique to the flexible devices of interest are still perceptible. In Section 8.3, the proposed approaches are tested and evaluated on smart meter data from the Costa Rican sub-metering study. Note that measurement data at the device level are not leveraged during the detection and disaggregation processes but only used as ground truth for the evaluation. The sensitivity of the different approaches to temporal resolutions between 1 minute and 30 minutes is analyzed. In Section 8.4, the proposed approaches are applied to

---

[1] The notion of standard smart meter data refers to active and possibly reactive power measurements at the level of end-users with a temporal resolution between 1 and 30 minutes, excluding sub-metering data.

70 households in order to obtain a high-resolution estimation of the share of flexible consumption at an aggregate level. Finally, Section 8.5 concludes the chapter and discusses future work.

8.2 METHODOLOGY

Most TCLs are influenced by the outside temperature. This is, for example, the case of heat pumps and space heaters that are highly active in winter and barely used in summer. This characteristic can be exploited by disaggregation approaches to compare long periods with high and low TCL activity and distinguish temperature-sensitive loads from the non-flexible base load [147, 340]. However, this is not applicable to cold appliances and electric water heaters whose energy consumption is relatively constant throughout the year. Regarding the latter, the authors in [158] mention a difference in hot water demand of only 20% between the coldest and warmest days of the year in Europe. Hence, the disaggregation of cold appliances and electric water heaters should rely on different properties. In this work, a combination of features specific to each device is actually leveraged. Notably, cold appliances and water heaters are turned on and off almost instantaneously and consume a nearly constant active power during ON periods. As observed in Section 5.3.1, the rated power of cold appliances is relatively low, whereas water heaters are among the largest power consumers in the domestic sector. Besides, electric water heaters do not consume or produce any reactive power.

The proposed detection and disaggregation approaches leverage these specific features, which allows the devices of interest to be distinguished from other loads in the same house on the sole basis of standard smart meter data. The detection of cold appliance and water heater loads are first presented in Sections 8.2.1 and 8.2.2, respectively. The load detection approaches rely on the power histogram of the total house load, from which the rated power of the devices of interest can be detected. Subsequently, Sections 8.2.3 and 8.2.4 detail different approaches to isolate the respective individual device's load profiles. In this work, load detection is a necessary condition before disaggregation. Note that the entire process of both the detection and disaggregation steps are unsupervised, i.e., only power measurements at the house level are required.

8.2.1 *Detection of Cold Appliance Load*

Even if the presence of a refrigerator in a house is highly probable, this should be confirmed by the data. The load of cold appliances such as refrigerators is characterized by a cycle ON/OFF behavior with relatively constant rated power. Such behavior is triggered by a local hysteresis controller which aims to maintain the internal temperature within a certain band. The underlying assumption for load detection is that the cyclic ON/OFF behavior translates into two significant spikes in the power histogram of the total house load. This is justified by the fact that cold appliances are continuously in activity, also in the absence of other loads. In order to avoid the interference of other TCLs in the detection process, only power values up to 1 kW are considered. On the one hand, this limit is well above the usual residential refrigerator rated power (i.e., maximum 200 W [287]) and also accounts for a small and base load (e.g., sleep mode of electronic devices). On the other hand, the upper limit lies below the rated power of other typical TCLs that might also induce spikes in the higher portion of the total load histogram[2] (e.g., minimum 2 kW for water heaters and AC units [287]). The histogram in question consists of 50 bins with a width of 20W, which allows for precise recognition of spikes. On this basis, a spike is defined as a bin whose two neighboring bins have a lower frequency of power values. A cold appliance is said to be detected if the two largest spikes are at least 25% higher than the remaining spikes. This criterion is defined based on a preliminary study. Besides, the difference in power between the spikes corresponds to the estimated rated power of the cold appliance.

Figure 8.1 illustrates the total active power histogram at different temporal resolutions of an example house with a refrigerator. Two significant spikes are visible (i.e., at 40W and 100W), especially at higher temporal resolution. The second spike tends to disappear with decreasing temporal resolution. This is explained by the fact that the duration of the refrigerator's ON period (i.e., 34 minutes on average in this example) gets too close to the temporal resolution and is affected by the smoothing effect detailed in Section 5.3. The right subplots of Figure 8.1 illustrate the histogram of the exact same house, where the refrigerator load has been subtracted. Only one significant spike is left, supporting the assumption that two significant spikes under 1 kW indicate the presence of a cold appliance.

---

[2] In this chapter, the total load refers to the load at the house level.

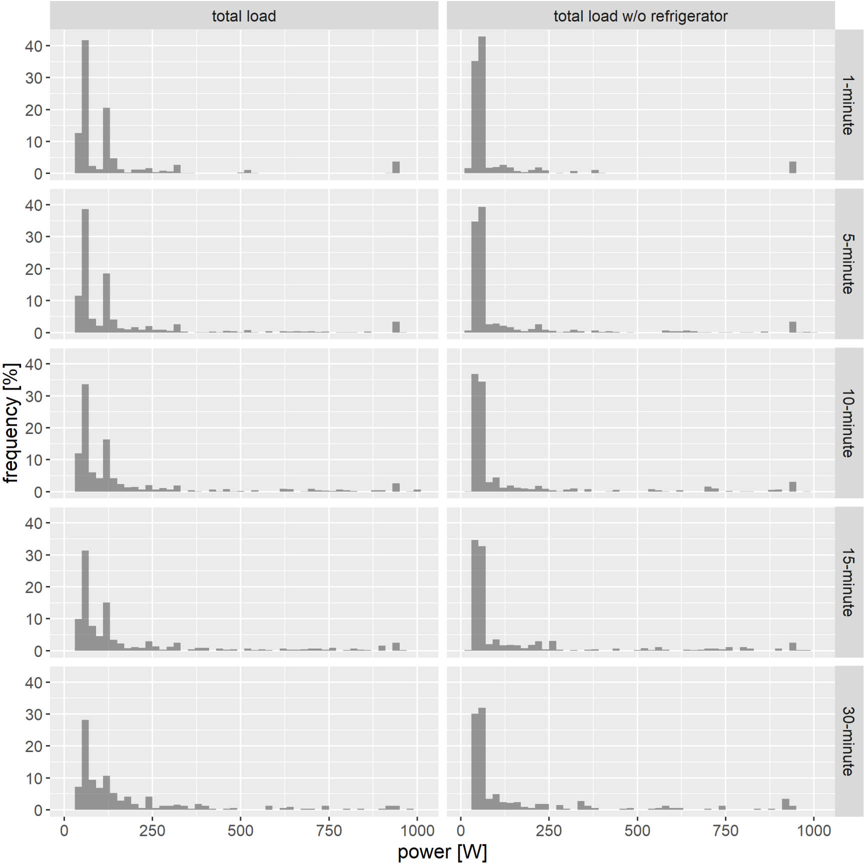


FIGURE 8.1: Active power histogram of an example house (i.e., house 6 in Table 8.1) with and without refrigerator.

8.2.2 *Detection of Water Heater Load*

The active power consumption profile of water heaters is characterized by large spikes of the same amplitude, corresponding to their rated power. The frequency and duration of these consumption spikes depend on the hot water demand and the thermal losses of the water tank. In general, they occur several times a day in order to maintain the water temperature in the tank within a certain range. Hence, the presence of electric water heaters is assumed to be visible in the upper part of the active power histogram of

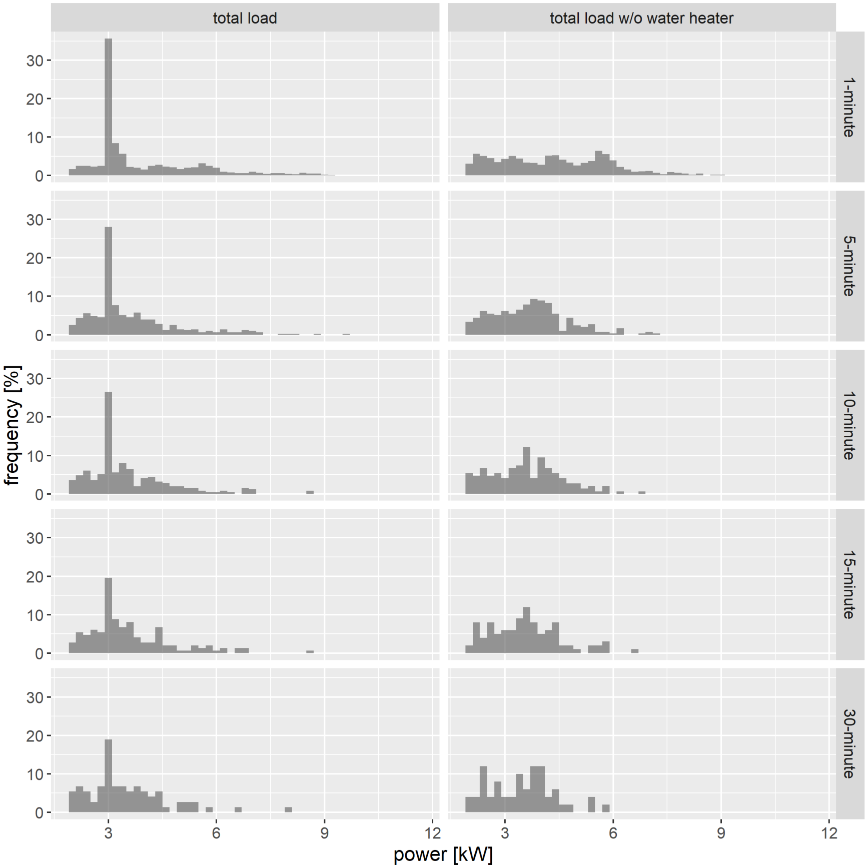


FIGURE 8.2: Histogram of the total active power consumption at various temporal resolutions in an example house (i.e., house 3 in Table 8.3) with and without a water heater.

the total house load. This is illustrated by Figure 8.2 for an example house at various temporal resolutions. Indeed, one of the bins is associated with a substantially higher frequency, which does not appear when the water heater load is subtracted from the total house load. Note also that the relative frequency of this bin diminishes with decreasing temporal resolution. This is again due to the smoothing effect of lower temporal resolutions on smart meter data, as detailed in Section 5.3. In order to filter out the activity of other TCLs with lower rated power (e.g., air conditioners and refrigerators), only power values higher than 2 kW are considered in the histogram. In this work, the width of the histogram bins is fixed to 200 W. A water heater is said to be detected if the power histogram contains at least one outlier bin, i.e., the frequency associated with at least one bin exceeds the commonly used threshold for outliers [238]:

$$d_i^{\text{bin}} >= d_{Q3}^{\text{bin}} + 1.5 \cdot (d_{Q3}^{\text{bin}} - d_{Q1}^{\text{bin}}), \tag{8.1}$$

where d_i^{bin} is the frequency associated to bin i in the total load histogram above 2 kW, and d_{Q1}^{bin} and d_{Q3}^{bin} are the first and third quartiles (i.e., 25^{th} and 75^{th} percentiles) among the different frequency values, respectively. At the difference of the criterion for the detection of cold appliance loads, there is no restriction on the number of outlier bins for the detection of water heater loads. The rated active power of the detected water heaters is estimated as the difference between the power of the highest outlier bin above 2 kW and the power of the first spike in the histogram below 2 kW. The latter corresponds to the base load. For the example house of Figure 8.2, the rated power of 3 kW can be correctly estimated.

8.2.3 *Disaggregation of Cold Appliance Load*

This section proposes two disaggregation approaches for cold appliance loads, assuming that the load detection step has been conclusive. The first approach distinguishes time periods of different activity levels in the total house load profile and learns the characteristics of the cold appliance from the so-called low-activity sections. The second approach is meant as an alternative in the case the distinction between periods of different activity levels is not feasible. It is based on the detection of active power jumps in the total load profile.

8.2.3.1 *Low-Activity-Based Approach*

The first proposed approach relies on the notions of “low-activity” and “high-activity” sections. Low-activity (LA) sections represent the portions of the time series where only the cold appliance and a potential small base load are active. In contrast, the activity of other loads is assumed to occur during the high-activity sections. The main idea is to understand the pattern of the cold appliance load in the LA sections and apply this knowledge in the HA sections for the purpose of disaggregation. More specifically, a LA section is defined as a time period where all values are lower than a certain threshold. In this case, the threshold corresponds to the power value of the so-called tail bin of the second significant spike in the total load histogram. The tail bin is defined as the last bin after the second significant spike whose difference in frequency with the previous bin is negative. Indeed, the variations in the measurements of the cold appliance's rated power (plus a potential base load) affect a few bins following the second significant spike, which is noticeable in Figure 8.1. AS safeguard, the threshold is nevertheless limited to 50% higher than the power value of the second significant spike. In addition, the cycle duration of cold appliances is generally lower than two hours according to Table 8.1 and the authors in [287, 355]. Hence, a minimum duration of two hours is required to ensure the isolation of at least one full ON/OFF cycle per LA section without the interference of other loads.

For each LA section, the disaggregated load is defined as the total load reduced by the corresponding minimum value:

$$p^{\mathrm{LA}}_{\mathrm{dis},t} = p^{\mathrm{LA}}_{\mathrm{tot},t} - \min_{j\in[1,T_{\mathrm{LA}}]} p^{\mathrm{LA}}_{\mathrm{tot},j}, \quad \forall t \in [1, T_{\mathrm{LA}}], \tag{8.2}$$

where $p^{\mathrm{LA}}_{\mathrm{dis},t}$ and $p^{\mathrm{LA}}_{\mathrm{tot},t}$ are the disaggregated and total house load at time t of the LA section under consideration, respectively. In addition, T_{LA} is the total number of time steps in this LA section. Following that, five representative LA sections are selected. As selection criterion, the features of their ON/OFF cycles (i.e., the power consumption and the time duration of the ON and OFF periods) must be the closest to the corresponding median values over all LA sections.

Subsequently, the disaggregation in each HA section is based on the disaggregated load in the representative LA sections and occurs in multiple steps:

1. The total load profile is modified in an iterative manner in order to improve the performance of the following steps. Concretely, the power jump (or power difference) between two consecutive time steps is limited to the largest jump observed in the representative LA sections while ensuring non-negative values:

$$\text{init} : p^{\mathrm{HA}}_{\mathrm{a}} = p^{\mathrm{HA}}_{\mathrm{tot}}, \tag{8.3a}$$

$$\text{iter} : p^{\mathrm{HA}}_{\mathrm{a},t:T_{\mathrm{HA}}} = p^{\mathrm{HA}}_{\mathrm{a},t:T_{\mathrm{HA}}} - \delta_i, \quad t = 2, \ldots, T_{\mathrm{HA}}, \tag{8.3b}$$

$$\text{with}\, \delta_i = \begin{cases} \max\left(0, \Delta p^{\mathrm{HA}}_{\mathrm{a},t} - \Delta p^{\mathrm{LA}}_{\mathrm{dis}}\right), \text{ if } \Delta p^{\mathrm{HA}}_{\mathrm{a},t} \geqslant 0, \\ \max\left(0, \Delta p^{\mathrm{HA}}_{\mathrm{a},t} + \Delta p^{\mathrm{LA}}_{\mathrm{dis}}\right), \text{ if } \Delta p^{\mathrm{HA}}_{\mathrm{a},t} < 0, \end{cases}$$

where $p^{\mathrm{HA}}_{\mathrm{a}}$ and $p^{\mathrm{HA}}_{\mathrm{tot}}$ are the iterated and total load profiles in the HA section under consideration, respectively. In addition, $\Delta p^{\mathrm{HA}}_{\mathrm{a},t} =$

$p_{\mathrm{a},t}^{\mathrm{HA}} - p_{\mathrm{a},t-1}^{\mathrm{HA}}$, $\Delta p_{\mathrm{dis}}^{\mathrm{LA}}$ is the largest absolute power difference between two consecutive time steps in the representative LA sections, T_{HA} is the number of time steps in this HA section, and subscript $t : T_{\mathrm{HA}}$ refers to time steps t to T_{HA} of the corresponding profile.

2. Each of the representative LA section is extended (i.e., duplicated by respecting the ON/OFF cycle features) to be longer the HA section under consideration by at least the length of the LA section. For example, an extended LA section should have at least $20 + 6 = 26$ time steps if the HA section and the original LA section consist of 20 and 6 time steps, respectively. Following that, The Pearson's correlation between the extended LA sections and the modified load in the HA section is computed for each possible lag. The load of the extended representative LA section with the lag leading to the highest correlation is selected as the best-fitting model:

$$p_{\mathrm{b}}^{\mathrm{HA}} = \underset{p_{\mathrm{lag}}^{\mathrm{LA}}}{\operatorname{argmax}} \operatorname{cor}\left(p_{\mathrm{a}}^{\mathrm{HA}}, p_{\mathrm{lag}}^{\mathrm{LA}}\right), \tag{8.4}$$

where $p_{\mathrm{b}}^{\mathrm{HA}}$ is an intermediate load profile and $p_{\mathrm{lag}}^{\mathrm{LA}}$ is an extended representative LA section with a certain lag.

3. The disaggregated load in the HA section is defined as the minimum at each time step between the best-fitting model and the original load:

$$p_{\mathrm{dis},t}^{\mathrm{HA}} = \min\left(p_{\mathrm{b},t}^{\mathrm{HA}}, p_{\mathrm{tot},t}^{\mathrm{HA}}\right), \quad \forall t \in [1, T_{\mathrm{HA}}], \tag{8.5}$$

where $p_{\mathrm{dis},t}^{\mathrm{HA}}$ is the disaggregated load at time t of the HA section under consideration.

To summarize, the total disaggregated load consists of $p_{\mathrm{dis}}^{\mathrm{LA}}$ and $p_{\mathrm{dis}}^{\mathrm{HA}}$ in LA and HA sections, respectively. The disaggregation process of a HA section is illustrated in Figure 8.3. Based on this approach, the disaggregated load looks like an actual cold appliance load, even in high-activity periods. Note that this approach is not applicable if other loads are continuously active, which prevents the isolation of LA sections.

8.2.3.2 *Jump-Based Approach*

In order to disaggregate the cold appliance load from any house's load profile, even if satisfactory LA sections are not visible, a second approach is proposed. In this case, power variations of the disaggregated load are constrained

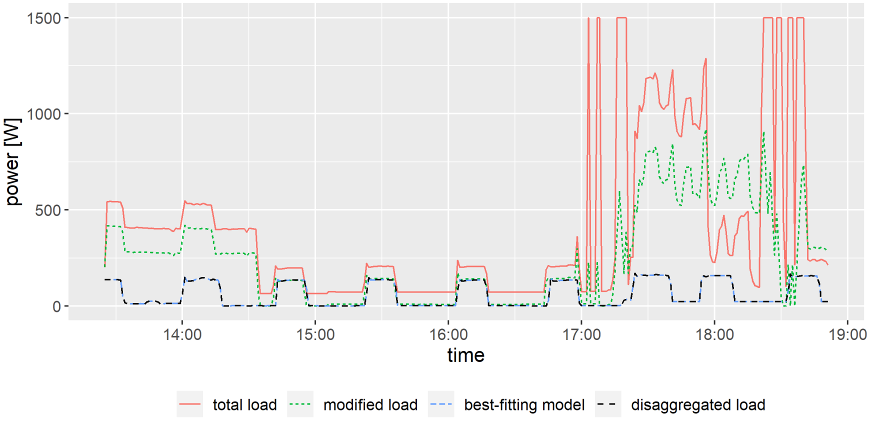


Figure 8.3: Disaggregation process of a refrigerator load according to the low-activity-based approach based on 1-minute resolution data. For visualization purposes, power values are capped at 1500W. The illustrated section corresponds to the squared area in Figure 8.7.

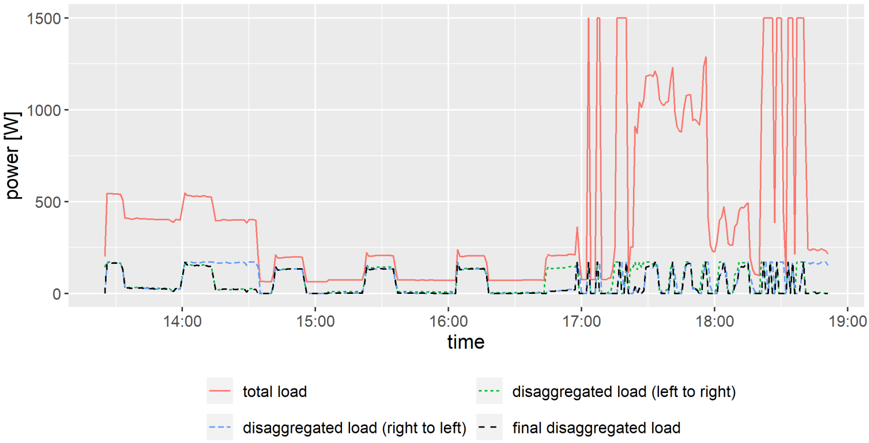


Figure 8.4: Disaggregation process of a refrigerator load according to the jump-based approach based on 1-minute resolution data. For visualization purposes, power values are capped at 1500W. The illustrated section corresponds to the squared area in Figure 8.7.

within the range of the estimated rated power of the cold appliance. The disaggregation process is given as follows:

1. The original load profile is modified in an iterative manner such that the jump in power between two time steps does not leave a predefined band. This step is carried out:
    a) from left to right:

$$\begin{aligned} &\text{init} : p_{\text{a}}^{\text{it}} = p_{\text{tot}}, && (8.6\text{a}) \\ &\text{iter} : p_{\text{a},t:T}^{\text{it}} = p_{\text{a},t:T}^{\text{it}} - \lambda_i, \quad t = 1, \ldots, T, && (8.6\text{b}) \\ &\text{with } \lambda_i = \begin{cases} p_{\text{a},t}^{\text{it}} - p_{\text{lim}}, & \text{if } p_{\text{a},t}^{\text{it}} > p_{\text{lim}} \\ p_{\text{a},t}^{\text{it}}, & \text{if } p_{\text{a},t}^{\text{it}} < 0 \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

    where p_{a}^{it} and p_{tot} are the iterated and total load profiles, respectively. In addition, T is the total number of time steps, and subscript $t : T$ refers to time steps t to T of the corresponding profile. In this way, the iterated profile is contrained between the power values of 0 W and p_{lim} which is set to the power value of the second spike in the total load histogram plus a margin of 20 W. This margin account for small variations in the power of the refrigerator (plus a potential base load)[3].

    b) from right to left:

$$\begin{aligned} &\text{init} : p_{\text{b}}^{\text{it}} = p_{\text{tot}}, && (8.7\text{a}) \\ &\text{iter} : p_{\text{b},1:t}^{\text{it}} = p_{\text{b},1:t}^{\text{it}} - \lambda_i, \quad t = T, \ldots, 1, && (8.7\text{b}) \\ &\text{with } \lambda_i = \begin{cases} p_{\text{b},t}^{\text{it}} - p_{\text{lim}}, & \text{if } p_{\text{b},t}^{\text{it}} > p_{\text{lim}} \\ p_{\text{b},t}^{\text{it}}, & \text{if } p_{\text{b},t}^{\text{it}} < 0 \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

    where the different variables and parameters are defined as in Equation (1a).

2. The disaggregated load profile is finally defined as the minimum between the profiles resulting from steps 1a and 1b at each time step:

$$p_{\text{dis},t}^{\text{jump}} = \min\left(p_{\text{a},t}^{\text{it}}, p_{\text{b},t}^{\text{it}}\right), \quad \forall t \in [1, T], \tag{8.8}$$

---

[3] The exact value of p_{lim} does not significantly impact the disaggregation outcome as long as it is not too large.

where $p_{\text{dis},t}^{\text{jump}}$ is the disaggregated load at time t.

The disaggregation process according to the jump-based approach is illustrated in Figure 8.4. For comparison purposes, the same section as in Figure 8.3 is selected. The shape of the final disaggregated load is highly sensitive to the variations in the total load and particularly volatile in periods with high activity.

8.2.4 *Disaggregation of Water Heater Load*

The proposed disaggregation method for the water heater load leverages active power measurements in the first step and, if available, reactive power measurements in the second step. The core idea of the disaggregation process relies on the detection of large jumps in the power profiles. This concept is illustrated in Figure 8.5 for an example house with 10-minute resolution data. The first subplot represents the total active power profile of the house, from which upward and downward jumps in the same range as the estimated rated power are identified. In the second subplot, upward and downward jumps are translated into values of 1 and -1, respectively. Note that the transition between the ON and OFF status of a water heater takes place almost instantaneously, which is much faster than the temporal resolution of the data. Due to the temporal averaging effect, the power value measured for the water heater at the time step when the transition occurs appears to be a weighted average between zero and the rated power. Hence, a jump considers the difference in power over three consecutive time steps. If available, the same procedure is applied to the reactive power profile, as illustrated in the third and fourth subplots of Figure 8.5. In order to account for small variations, the threshold for the identification of large jumps in active power is set 10% below the rated power estimated from the total active power histogram as presented in Sec. 8.2.2. Analogously, the threshold for large reactive power jumps is set 10% below the most frequent absolute reactive power value observed in the histogram (bin width of 20 Var) of the total reactive power[4].

In a further stage, the identified jumps in the total active power profile are cleaned up to ensure an alternation of upward and downward jumps. Hence, the ON periods of the water heater are defined between upward and downward jumps, and the OFF periods are defined between downward and upward jumps. Moreover, only ON periods with a duration lower than 2

---

[4] The exact value of the power thresholds only marginally impacts the disaggregation outcome as long as they are slightly lower than the estimated rated powers.

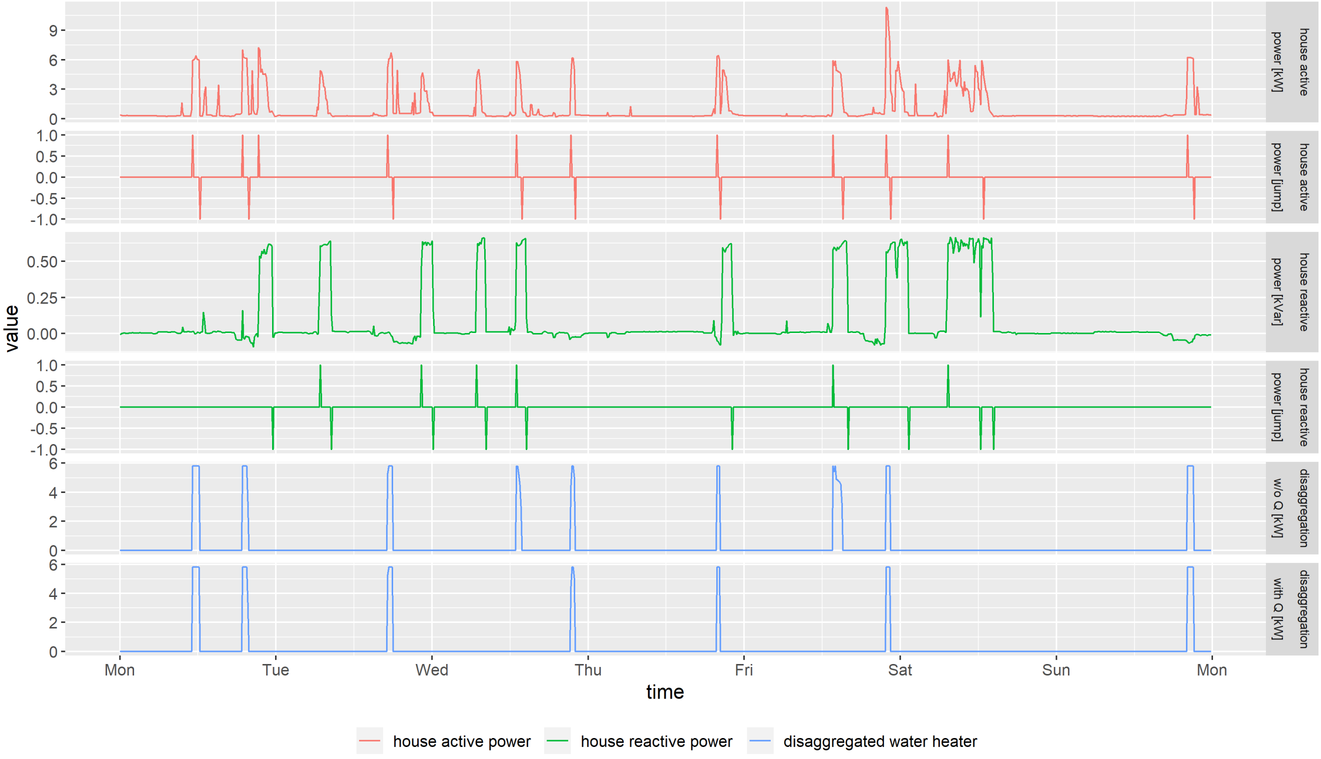


FIGURE 8.5: Disaggregation process of a water heater load in an example house (i.e., house 1 in Table 8.3) based on 10-minute resolution active and reactive power measurements over one week.

hours are retained, which is a realistic upper limit for typical water heaters according to Tables 8.2 and 8.3, and to the authors in [287]. The resulting active load of the water heater is zero during OFF periods and is equal to the minimum between the estimated rated power and the total active power during ON periods:

$$p_t^{\mathrm{WH}} = \begin{cases} 0, & \text{if } \ t \in \text{OFF period}, \\ \min\left(\hat{p}^{\mathrm{WH}}, p_t^{\mathrm{tot}}\right), & \text{if } \ t \in \text{ON period}, \end{cases} \qquad \forall t \in [1, T], \tag{8.9}$$

where p_t^{WH} is the estimated active power of the water heater at time t, $\hat{p}^{\mathrm{WH}}$ is the estimated rated power of the water heater, p_t^{tot} is the active power of the total house at time t, and T is the number of time steps. As illustrated in the fifth subplot of Figure 8.5, only part of the visible spikes in total active power are associated with the water heater activities. The remaining spikes either have a too low maximum power, last longer than two hours, or require more than two time steps to reach their maximum. According to the proposed disaggregation approach, they likely correspond to the activity of other electric devices.

Finally, reactive power measurements can be leveraged to enhance the precision of the disaggregation process. Indeed, electric water heaters are purely resistive devices (i.e., consuming only active power), in contrast to most other electric devices which are usually equipped with induction motors. This is confirmed by Table 6.1. Hence, the simultaneous detection of a jump in both active and reactive power profiles certainly reflects the activity of an inductive (or capacitive) electric device and cannot be attributed to a water heater. In this way, the initial guess for the water heater power profile can be further enhanced. More precisely, wrongly estimated ON periods are filtered out by the processing of reactive power measurements. The comparison of the two last subplots of Figure 8.5 indicates that two ON periods are filtered out. They correspond to the activity of another electric device with a similar active power level but also reactive power consumption.

8.3 PERFORMANCE EVALUATION

The performance of the proposed detection and disaggregation approaches is evaluated on real-world data sets with sub-metering. In the following, Section 8.3.1 presents the data sets, and Section 8.3.1 defines the two types of evaluation metrics used in this study. The evaluation itself is given in Sections 8.2.3 and 8.2.4 for the cold appliances and water heaters, respectively.

8.3.1 *Data Set*

The measurement data for testing the proposed load disaggregation approach for cold appliances are obtained from the Costa Rican sub-metering study presented in Section 3.2.3. Out of the 70 houses under consideration, 14 examples consist of clean and reliable active power measurements of both the total load and the refrigerator load are available. The remaining houses probably also possess a refrigerator, but the power consumption has not been recorded separately, or the data quality is not sufficient. The preparation of the sub-metering data set is detailed in Section 5.2.1. Table 8.1 details the specifications of the refrigerators in question. Especially, their rated power is relatively low in comparison with other domestic appliances, ranging from 100 W to 200 W. Nevertheless, refrigerators consume between 11% and 51% of the total household energy consumption, which is non-negligible in terms of flexibility potential. Note also an average ON duration between 11 and 38 minutes, which influences the disaggregation ability.

Moreover, 24 good-quality load profiles of water heaters have been individually recorded in the Costa Rican sub-metering study. Their features are summarized in Table 8.2. Although water heaters consume power during a relatively short period, they represent 25% on average of the total house's energy consumption due to their relatively high rated power. The duration of their ON periods largely varies across the different examples, which impacts the success of the disaggregation approach. Unfortunately, only active power has been recorded in this data set. To the best of the author's knowledge, no data set with active and reactive power measurements of both the total house load and the water heater load is publicly available. In order to evaluate the benefit of reactive power for disaggregation purposes, own power measurements have been conducted over two to four weeks in three different households of the City of San José, Costa Rica. Table 8.3 details the specifications of the water heaters in each house, which is consistent with the features of the 24 previously mentioned water heaters.

All measurement data are originally recorded with 1-minute granularity. In order to reflect lower temporal resolutions, power values are appropriately modified according to Equation (3.10). Typical SM resolutions of 5, 10, 15, and 30 minutes are taken into account. As noted in Chapter 5, some countries opted for a temporal resolution of 60 minutes that is however not considered in this case study. In fact, the resulting residential load profiles get particularly smooth, which hinders the detection and the proper disaggregation of the cold appliance and water heater loads.

House	Rated power [W]	Mean ON | OFF duration [min]	Share of energy [%]
1	100	38 | 24	13
2	135	16 | 25	24
3	115	25 | 54	16
4	150	18 | 25	11
5	115	20 | 44	15
6	100	34 | 59	14
7	120	20 | 45	15
8	120	21 | 57	18
9	150	27 | 41	14
10	160	11 | 20	20
11	140	16 | 43	45
12	150	17 | 19	20
13	150	17 | 19	25
14	200	14 | 37	51

TABLE 8.1: Specifications of 14 refrigerators in the Costa Rican sub-metering data set.

Measure	Rated power [kW]	Mean ON | OFF duration [min]	Share of energy [%]
Average	4.8	8.7 | 384	25
Standard deviation	2.1	5.9 | 270	13.6
Minimum	2.9	3.1 | 87.6	7
Maximum	10.3	32.1 | 1396	63

TABLE 8.2: Statistics over 24 water heaters in the Costa Rican sub-metering data set.

8.3.2 *Evaluation Metrics*

The performance of the disaggregation approaches is evaluated based on two different types of metrics. On the one hand, error metrics focus on the

House	Rated power [kW]	Mean ON | OFF duration [min]	Share of energy [%]
1	6	26 | 628	27
2	6	4 | 385	17
3	3	30 | 410	29

TABLE 8.3: Specifications of 3 water heaters used in the evaluation.

point-wise mismatch between the true and the estimated power values. On the other hand, classification metrics assess whether the ON and OFF periods of the load are estimated at the correct time steps.

8.3.2.1 *Error Metrics*

Commonly used in the forecasting and disaggregation literature, the Mean Error (ME), the Mean Absolute Error (MAE), and the Normalized Mean Absolute Error (NMAE) consider the point-wise difference between two time series. They are defined as follows:

$$\mathrm{ME} = \frac{1}{T} \cdot \sum_{t=1}^{T} \left(\hat{y}_t - y_t\right), \tag{8.10a}$$

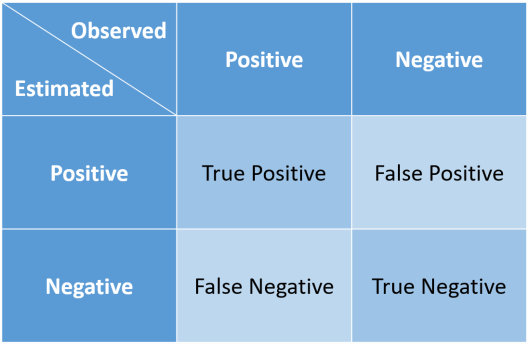
$$\mathrm{MAE} = \frac{1}{T} \cdot \sum_{t=1}^{T} |\hat{y}_t - y_t|, \tag{8.10b}$$

$$\mathrm{NMAE} = 100\% \cdot \frac{1}{T \cdot y_r} \cdot \sum_{t=1}^{T} |\hat{y}_t - y_t|, \tag{8.10c}$$

where y_t and $\hat{y}_t$ are the true and estimated values at time t, respectively. In addition, y_r is the actual rated value of the device under consideration, and T is the total number of time steps. While the ME makes the distinction between an overestimation and an underestimation of the actual values, the MAE gives the estimation error in absolute terms. Besides, the NMAE allows for fair comparison between multiple houses and devices of various sizes. For the error metrics, values close to zero indicate a good performance.

8.3.2.2 *Classification Metrics*

The disaggregation of a TCL can also be seen as the estimation of ON (positive) and OFF (negative) events at each time step, regardless of the



Observed / Estimated	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

FIGURE 8.6: Typical confusion matrix used as basis for classification metrics.

power magnitude. In this study, the load of the device of interest at a certain time step is considered as a positive event when it exceeds one third of the actual rated power, and as negative event otherwise. On this basis, disaggregation models can be evaluated like a classification problem. Hence, following classification metrics are considered:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}, \tag{8.11a}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}, \tag{8.11b}$$

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{8.11c}$$

where true positives are correctly estimated ON events, false positives are incorrectly estimated ON events, and false negatives are incorrectly estimated OFF events. Together with the true negatives (i.e., correctly estimated OFF events), their numbers build a so-called confusion matrix which is illustrated in Figure 8.6. On this basis, the precision is the ratio of correctly estimated ON events over all estimated ON events, the recall is the ratio of correctly estimated ON events over all actual ON events, and the F1-score is the harmonic mean between the precision and the recall. For the classification metrics, the higher the value, the better the performance.

8.3.3 *Disaggregation of Cold Appliance Load*

In order to get a first intuition of the disaggregated refrigerator load, Figure 8.7 illustrates the outcomes of both disaggregation approaches at different

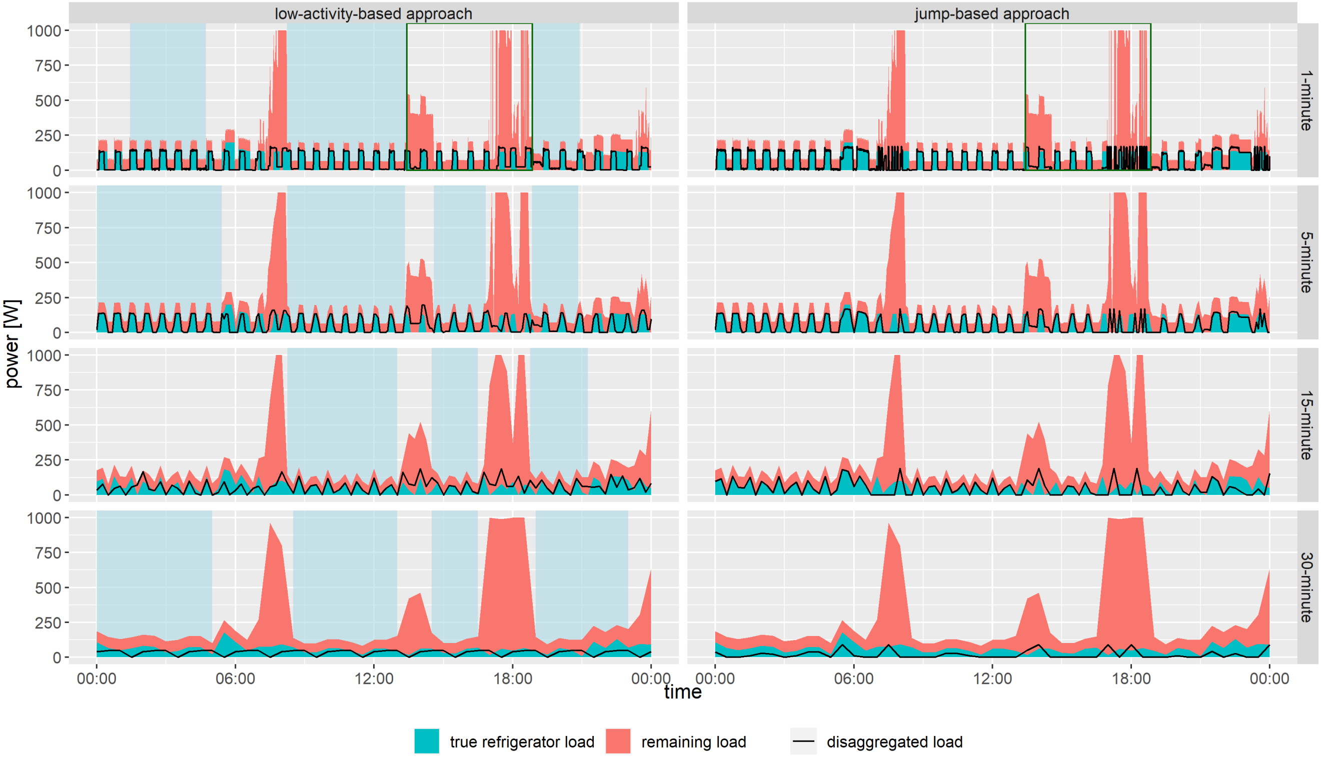


FIGURE 8.7: Disaggregation of the refrigerator load in an example house (i.e., house 2 in Table 8.1) over one day at different temporal resolutions according to both proposed approaches. For visualization purposes, power values are capped at 1000W. Blue-shaded areas are low-activity periods.

temporal resolutions for an example house. The disaggregation process in the squared areas is detailed in Figures 8.3 and 8.4 for low-activity-based and jump-based approaches, respectively. The shape of the true refrigerator load is highly affected when the temporal resolution decreases. The temporal averaging effect is all the more visible from a resolution of 15 minutes on, which is close to the mean duration of the refrigerator's ON and OFF periods (i.e., 16 and 25 minutes). The success of both disaggregations appears to suffer from a drop in temporal granularity. Nevertheless, the disaggregated load in LA sections matches remarkably with the true refrigerator load at high resolution. Whenever other loads are present, the jump-based approach estimates only a small portion of the refrigerator load. In contrast, the low-activity-based approach reproduces the cyclic behavior but sometimes misestimates the duration of the ON and OFF periods or the exact time of occurrence.

In a second step, the disaggregation performance is assessed over the 14 examples of houses with sub-metered refrigerator load[5]. Table 8.4 gives the number of houses for which disaggregation is possible. Since the jump-based approach can always be applied as long as a refrigerator load has been detected, the corresponding numbers also reflect the numbers of successful load detection[6]. The low-activity-based approach additionally requires the presence of LA periods of sufficient quality to be applied. Hence, the ratio of successful disaggregations is generally lower than for the competing approach. Moreover, the temporal resolution impacts the load detection ratio, which drops from 100% at a 1-minute resolution to 64% at a 30-minute resolution.

Concerning the actual disaggregation performance, the mean error in Figure 8.8 indicates that the jump-based approach is more conservative and generally underestimates the refrigerator load. In contrast, the low-activity-based approach exhibits an average mean error close to zero. The two other error metrics show that the performance of the low-activity-based approach is relatively independent of the temporal resolution. Conversely, the jump-based approach experiences a drop in performance with decreasing temporal resolution. A larger variance in accuracy is also visible across the multiple examples. On average, the low-activity-based approach outperforms the jump-based approach from a 15-minute resolution on. Depending on the resolution and the example in question, the mean absolute error ranges from

---

[5] Remember that sub-metering data are not required in the disaggregation process but are only exploited for evaluation purposes.

[6] Note that the performance of the load detection approach is not assessed via the typical classification metrics since there is no guarantee of the absence of a refrigerator in the remaining households of the Costa Rican sub-metering study. Hence, a confusion matrix cannot be built.

Temporal resolution	Jump-based approach	Low-activity-based approach
1 minute	14/14 (100%)	10/14 (71%)
5 minutes	13/14 (93%)	9/14 (64%)
10 minutes	13/14 (93%)	11/14 (79%)
15 minutes	13/14 (93%)	9/14 (64%)
30 minutes	9/14 (64%)	8/14 (57%)

TABLE 8.4: Number of successful disaggregations of refrigerator load according to the temporal resolution and the disaggregation approach.

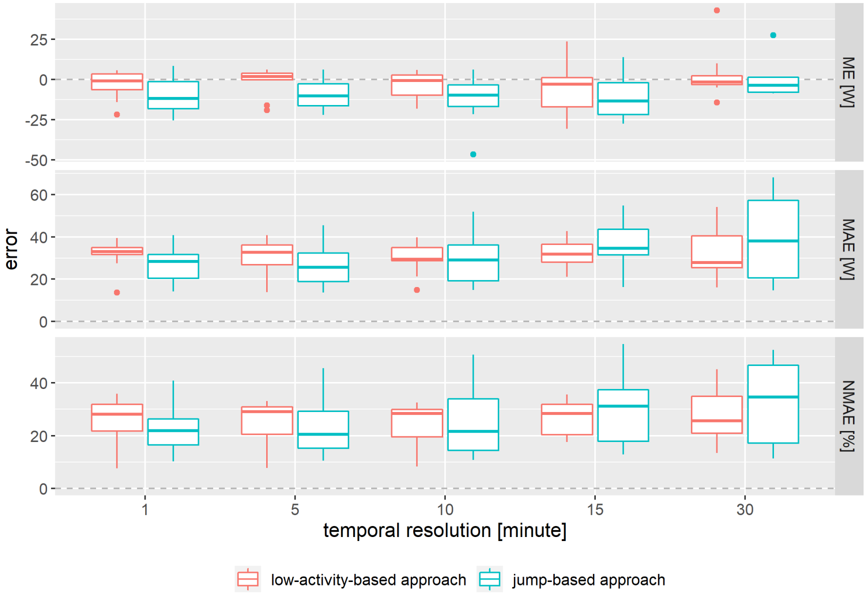


FIGURE 8.8: Performance evaluation of the refrigerator load disaggregation based on error metrics.

14 W 68 W with an average of 31 W. This corresponds to 7.5% to 55% with an average of 26% of the respective refrigerators' rated power (i.e., NMAE).

Finally, Figure 8.9 indicates that both the precision and the recall are negatively affected by a decreasing temporal resolution. The drop in perfor-

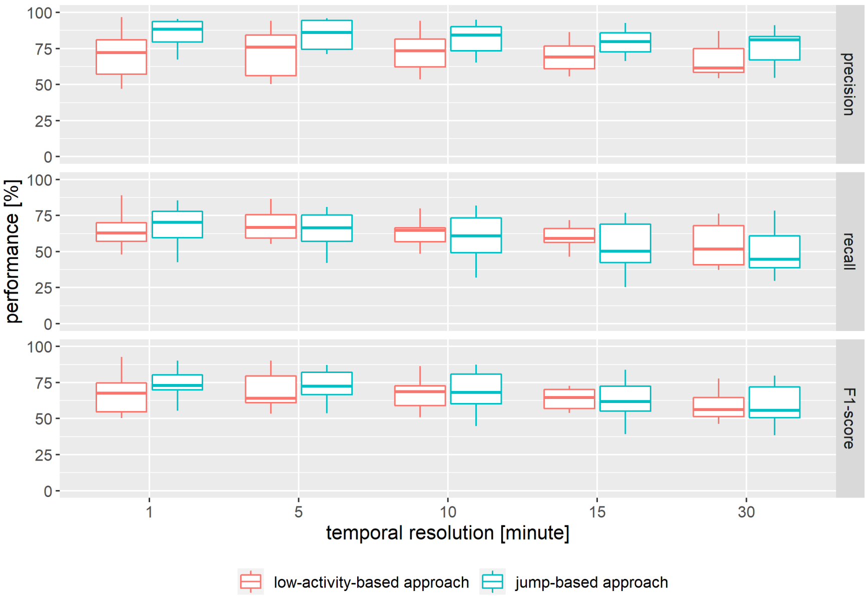


FIGURE 8.9: Performance evaluation of the refrigerator load disaggregation based on classification metrics.

mance can be explained by the misestimation of the actual refrigerator's rated power and the disappearance of the typical ON/OFF cycles. The jump-based approach is more precise than the low-activity-based approach but tends to miss more actual ON events from a 10-minute resolution on. The average F1-score goes from 68% and 75% at a 1-minute resolution to 58% and 59% at a 30-minute resolution for the low-activity-based and jump-based approaches, respectively. Although it is difficult to compare the performance on different data sets, still note that the average F1-score of the best-performing supervised classifier proposed in [349] amounts to 62% 32% and 20% at 1-, 5-, and 15-minute resolutions, respectively. Even if they are not supervised, the low-activity-based and jump-based approaches solely focus on the specific features of cold appliances, which leads to a very satisfying disaggregation performance accounting for the large uncertainty during periods of high activity.

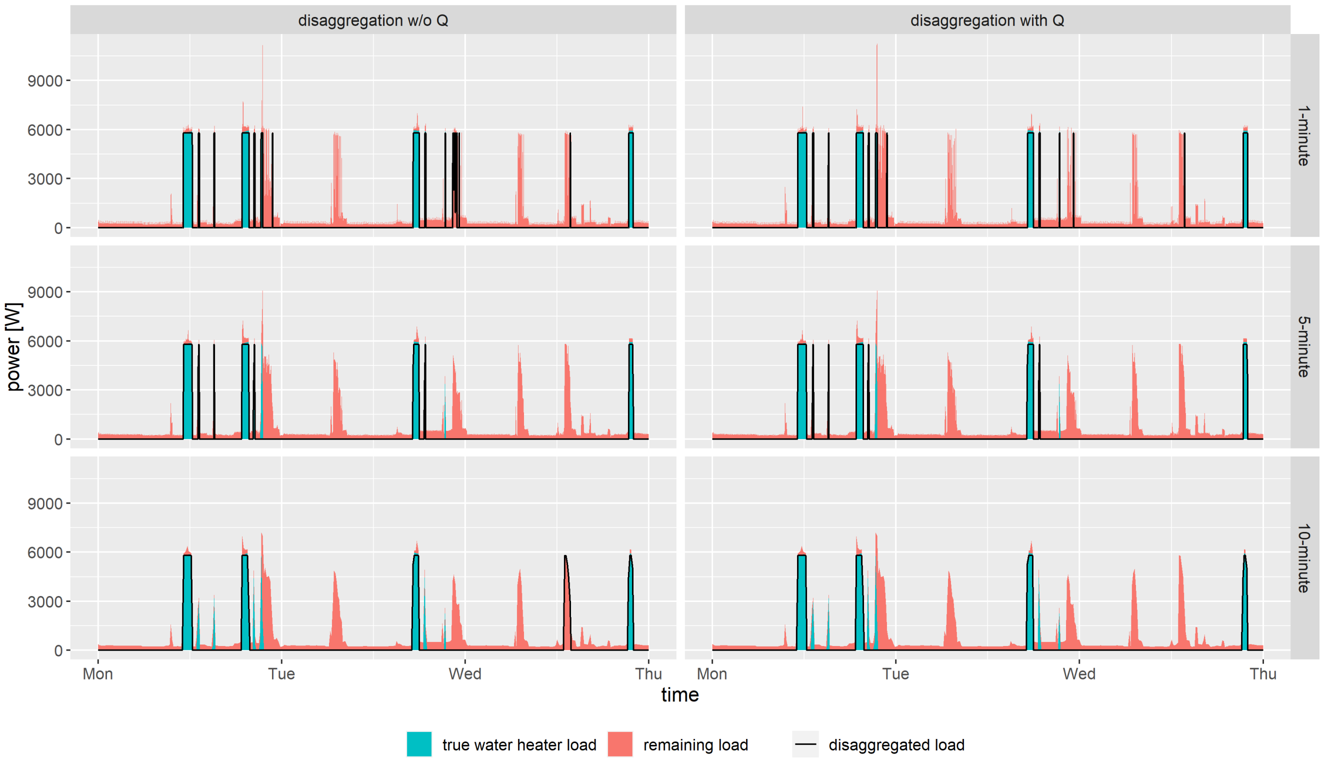


FIGURE 8.10: Disaggregation of the water heater load in an example house (i.e., house 1 in Table 8.3) over three days at different temporal resolutions, with and without the help of reactive power (Q).

8.3.4 *Disaggregation of Water Heater Load*

The disaggregation of the water heater load is evaluated in a similar manner. First, Figure 8.10 allows for an intuitive interpretation of the outcome. The true and estimated water heater load profiles are illustrated over the first three days of Figure 8.5. In addition to the water heater with a rated power of 6 kW, another electric device consumes active power with a similar magnitude, which challenges the water heater disaggregation process. As repeatedly observed, the temporal resolution impacts the shape of the total load profile and also alters the disaggregation of the water heater load. In this example, the water heater load cannot be detected anymore from At a 1-minute resolution, all water heater events are correctly detected, but the proposed approach misclassifies some events of the other large load which shows similar features as the water heater (false positive). The misclassification is however mitigated by the processing of reactive power measurements. At lower temporal resolutions, the magnitude of the narrow and successive consumption spikes induced by the other large load is partially reduced below the detection threshold related to the water heater (e.g., on Monday and Tuesday evening). This contributes to eliminating some false positives observed at a 1-minute resolution. At a 10-minute resolution, the proposed approach still detects water heater events with a long duration but misses narrower water heater events due to the smoothing effect (e.g., on Monday afternoon). In this case, the use of reactive power measurements also allows the proposed approach to dispose of a false positive on Wednesday afternoon.

Figures 8.11 and 8.12 illustrate the disaggregation performance for the three water heaters specified in Table 8.3. First, note that the water heater is not detected in house 1 with 15-minute data and in house 2 with 10- and 15-minute data. As previously explained, this is directly linked to the ON duration of the water heaters, which determines if the observed rated power is affected by the temporal averaging effect. Although the water heater of house 2 is still detected at a 5-minute resolution, the smoothing effect already impacts a considerable part of the ON periods that cannot be identified anymore, which translates into a substantial drop in precision and an increase of the absolute error metrics.

Furthermore, the presence of highly volatile loads with a similar (or higher) power amplitude as the water heater in houses 2 and 3 explains the relatively lower precision and higher error at 1-minute resolution. This is partially mitigated by the processing of reactive power measurements and is smoothed out at lower temporal resolutions. More generally, leveraging reactive power

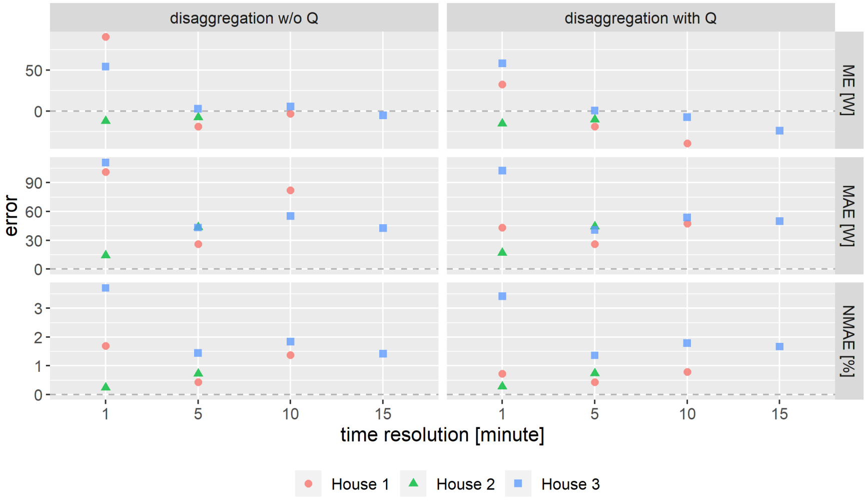


FIGURE 8.11: Performance evaluation of the water heater load disaggregation based on error metrics, accounting for the use of reactive power.

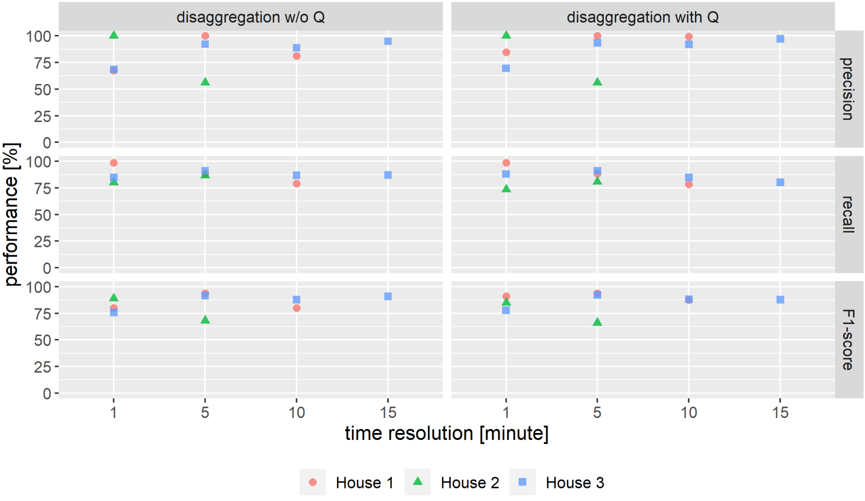
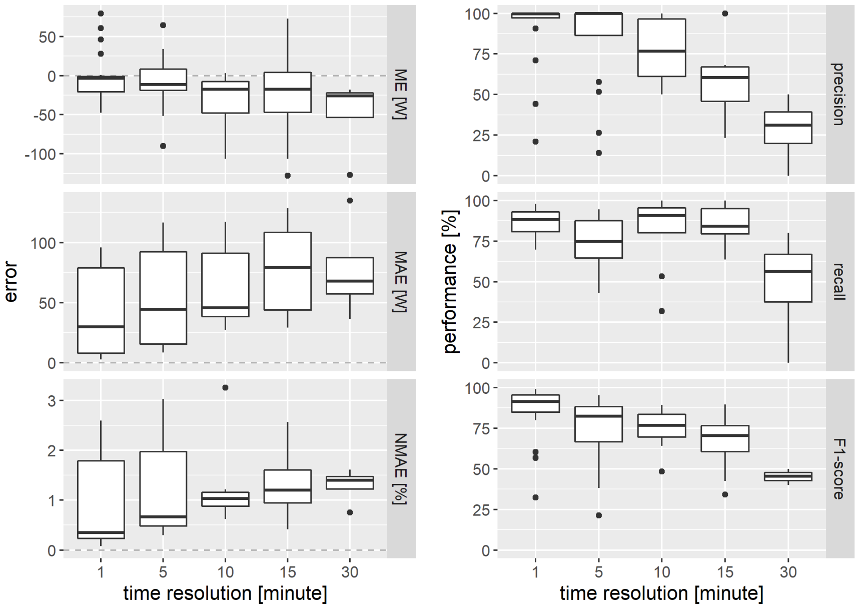


FIGURE 8.12: Performance evaluation of the water heater load disaggregation based on classification metrics, accounting for the use of reactive power.



(a) Error metrics (b) Classification metrics

FIGURE 8.13: Performance evaluation of the water heater load disaggregation on the Costa Rican sub-metering data set.

measurements during the disaggregation process typically increases the disaggregation precision but sometimes at the cost of a slightly lower recall. Indeed, in rare cases, the water heater is switched ON or OFF at the exact same time as another load with reactive power consumption. Consequently, by leveraging reactive power, the algorithm wrongly considers the jump in active power as a false positive.

In order to obtain a more representative picture of the disaggregation accuracy, the analysis is also performed on the 24 households with good-quality water heater load profiles from the sub-metering data set, however with the possibility to exploit reactive power. First of all, the load detection steps succeed for 20, 16, 10, 8, and 4 examples based on data with temporal resolutions of 1, 5, 10, 15, and 30 minutes, respectively. Figures 8.13a and 8.13b also confirm the high impact of the temporal averaging effect on the general performance. The mean error indicates a slight underestimation of the load at lower resolutions. A relatively large variance across the different examples appears in terms of point-wise error. The precision and the recall are drastically impacted from a resolution of 10 minutes and 30 minutes on, respectively. In general, note the excellent performance at a 1-minute resolution with median values of 41 W and 0.9% for the MAE and the NMAE, respectively. In terms of classification error, 1-minute resolution data allow for a mean precision, recall, and F1-score of 0.89, 0.86, and 0.85, respectively.

8.4 SHARE OF FLEXIBLE LOADS AT AN AGGREGATE LEVEL

Finally, the proposed detection and disaggregation approaches are applied to all 70 households in the Costa Rican sub-metering data set with a 1-minute resolution. In this study, the load of cold appliances is disaggregated via the jump-based approach since it is not restricted by the availability of low-activity sections. In addition, the water heater disaggregation approach relies only on active power measurement. It comes out that 49 water heaters and 65 cold appliances are detected and disaggregated, which seems sensible. In a minority of houses, it is however probable that the intense activity of other electric devices blurs the observation and detection of the appliances of interest. Note that the accuracy of the proposed approaches cannot be evaluated because of the lack of appropriate sub-metering data for all households. Nevertheless, this study first proves their scalability. In this case, the aggregated load is in the same order of magnitude as the load fed by a small MV/LV transformer. Since the approaches do not imply any machine learning and are purely based on data analytics, the time complexity and the computational resources are not a limiting factor to their application at the level of an entire distribution grid. On average, the disaggregation of each example in this case study lasts a couple of seconds. Furthermore, the high-resolution disaggregation of flexible loads provides precious insight into the demand response potential inherent to a given grid.

The top subplots of Figure 8.14 illustrate the average daily aggregate load profile of the disaggregated appliances on weekdays and at the weekend. Cold appliances are characterized by a relatively constant load of about 3 kW. In contrast, the aggregated water heater load exhibits a large variation over the day, where most of the consumption is concentrated in the morning. On weekdays, the water heater consumption peak occurs between 6:00 and 7:00, which corresponds to the time period when inhabitants are mostly expected to use hot water (e.g., shower, breakfast). At the weekend, this peak is displaced to the late morning. In both cases, the water heater consumption peak coincides with the total consumption peak. The share of the respective flexible loads with respect to the total load is illustrated in the bottom

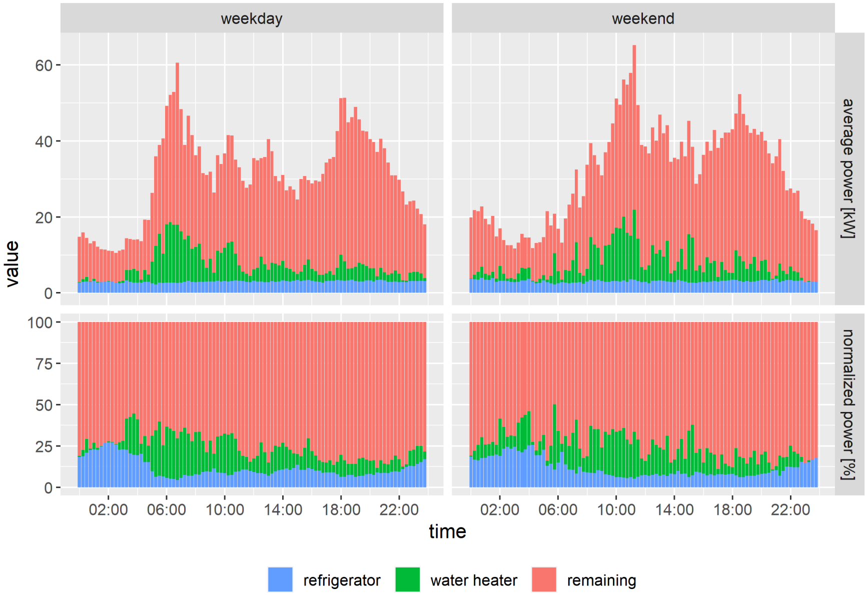


Figure 8.14: Average power consumption and share of disaggregated cold appliances and water heaters in the aggregation of the 70 household loads from the Costa Rican sub-metering data set, categorized by weekday and weekend.

subplots of Figure 8.14. Not surprisingly, the share of cold appliances reaches its maximum (i.e., 28%) at night, when the total load is at its minimum. The aggregated load of water heaters tends to follow the total load shape such that their contribution to the total load is somewhat spread over the day. Independently from the type of day, their consumption share reaches up to 35% in the early morning.

8.5 CONCLUSION

To summarize, this chapter presents multiple novel unsupervised approaches to detect and disaggregate the consumption profile of flexible TCLs from standard smart meter data. To the best of the author's knowledge, this has never been proposed in the literature for cold appliance and storage water heater loads. In fact, traditional NILM approaches make use of measurement data at temporal resolutions within the range of seconds and sometimes of sub-metering data, whose availability is an exception in common distribution grids. The proposed approaches only require power measurements with a typical resolution between 1 and 30 minutes gathered at the residential end-consumer level. It must nevertheless be noted that the success of detection and disaggregation approaches is negatively impacted by the temporal averaging effect occurring at lower resolutions.

Both load detection approaches rely on the total active power histogram, where the rated power of the device of interest is estimated. Subsequently, two different disaggregation algorithms are developed for cold appliances. Evaluated on 14 Costa Rican households with sub-metering data, a conservative jump-based approach performs particularly well at higher temporal resolutions. At 1- and 5-minute resolutions, it achieves an average NMAE of 21% and 22% (normalized by the rated power) and an average precision of 86% and 85%, respectively. The alternative approach learns the shape of the cold appliance load in so-called low-activity time periods without the interference of other devices and leads to the disaggregation of realistic load patterns in time periods of high activity. This low-activity-based approach appears to be less sensitive to the temporal resolution of the measurement data. Regarding the water heater, the proposed disaggregation algorithm detects the presence of large jumps in the total active power profile. It performs especially well at higher resolutions. Evaluated on 24 Costa Rican households with sub-metering data, the disaggregation approach achieves an average NMAE of 0.9% and 1.2%, and an average precision of 88% and 84% with temporal resolutions of 1 and 5 minutes, respectively. If available, reactive power measurements can enhance the precision of the disaggregation by eliminating false positives. The algorithm tends to be conservative, which is reflected by a generally higher precision than recall and a mean error slightly lower than zero.

Due to their relatively low computational complexity, the proposed approaches can easily be implemented on a wider scale. Notably, their application in a smart metered section of an LV grid enables to quantify the share of consumption coming from flexible devices with a high temporal and spatial resolution. In the presented case study with 70 households, it appears that cold appliances and water heaters account for a substantial portion of the total load. During the early morning consumption peak, water heaters are responsible for an estimated 35% of the total load. Even if their rated power is relatively low, refrigerators are continuously in activity and still contribute to an estimated 12% of the total load on average. This reflects a substantial flexibility potential for DR schemes which can be greatly enhanced by accurate estimation of the load profiles of individual devices. It must nevertheless be noted that cold appliances and water heaters cannot be entirely and permanently exploited for DR purposes. Hence, the share of their load does not directly reflect the actual flexibility potential. Future work shall focus on data-based flexibility estimation. This could rely on the data-based modeling of the internal bang-bang controller as proposed in [356] for the disaggregation of heat pump loads. Eventually, the design of online disaggregation algorithms based on real-time SM data streams would definitely support the operation of active distribution grids.

9

DETERMINISTIC AND PROBABILISTIC SHORT-TERM FORECASTING

> *Prediction is very difficult, especially if it's about the future.*
>
> — Nils Bohr

This chapter elaborates on the application of short-term forecasting in low-voltage grids. Short-term load forecasting becomes popular at this level due to the availability of smart meter measurements. Nevertheless, corresponding load profiles are characterized by much higher volatility and lower predictability than the load at higher spatial aggregation levels. In this context, traditional deterministic forecasting algorithms and standard metrics, which focus on point-wise accuracy, seem inadequate. In contrast, probabilistic approaches allow for a certain quantification of the large uncertainty inherent to low-voltage systems. Quantile forecasting can be leveraged to estimate in a probabilistic way the near-future system state, including bus voltages and line power flows. Among others, such short-term probabilistic state forecasting benefits the design of preventive voltage control schemes. This chapter discusses these different aspects in detail. It is based on [134] and [357] for the deterministic and probabilistic parts, respectively.

9.1 INTRODUCTION

As repeatedly mentioned over the course of this thesis, the digitalization of distribution grids opens up a large variety of new opportunities in terms of grid transparency to support operation and planning processes. For example, the authors in [4] review a large set of domains that are boosted by the wide-scale roll-out of smart meters. They mention the efficient integration of Distributed Energy Resources (DERs) and the management of flexible loads via demand response strategies as promising applications of smart meter data analytics. In fact, the success of these processes inevitably relies on good predictions, not only at an aggregate level but also down to the end-consumer level. Hence, an increasing number of deterministic models have been designed and proposed in the literature for load forecasting at the level of end-users or small aggregations of consumers [286, 358–363]. In the large majority of cases, they are not more than an adaptation of widely known algorithms traditionally applied at the transmission level and in wind forecasting, e.g., autoregressive models, artificial neural networks, regression trees, support vector regressions, k-nearest neighbors, fuzzy techniques, and hybrid methods [9]. The selection of appropriate features appears to be particularly relevant. Although some algorithms might be more efficient, it seems clear that the performance significantly drops in comparison with load forecasting at the substation or transmission level [123]. The behavior of individual end-users is inherently harder to predict than their aggregation. In addition, the actual definition of a good prediction at low aggregation levels is not straightforward anymore due to the high load volatility. The authors in [136, 364] question the validity of standard evaluation metrics (e.g., mean absolute percentage error, root mean square error) which purely rely on the point-wise accuracy. In this context, the general application and evaluation of deterministic forecasting algorithms in Low-Voltage (LV) grids merits closer scrutiny.

Besides deterministic algorithms, probabilistic forecasting approaches are gaining popularity in power systems since they account for the uncertainty related to a given quantity. Nevertheless, their application to distribution grids, and especially to the LV level, is particularly scarce and exclusively focuses on PV production and load forecasting. For example, the authors in [365] propose a quantile version of Convolutional Neural Network (CNN) for the day-ahead and intra-day prediction of aggregations of 10 consumers. Besides, the authors in [137] demonstrate the superiority of a proposed pinball loss guided Long-Short-Term Memory (LSTM) approach over the quantile versions of Regression Neural Network (RNN) and Gradient Boosting Regression Tree (GBRT) for the 30-minute-ahead probabilistic forecasting of individual loads. In addition to the traditional load, the increasing share of Photovoltaic (PV) systems and Electric Vehicles (EVs) adds uncertainty in current distribution grids and introduces new operational challenges to DSOs. On the one hand, PV systems synchronously inject active power into the LV grid according to the volatile solar irradiance, which can considerably increase the voltage in the system. Moreover, reverse power flows can damage the current transformer infrastructure that is traditionally not designed for such cases. On the other hand, EVs represent large stochastic loads which increase the loading of the system and potentially decrease the voltage during the charging phase. Since DSOs are responsible for the safe operation of their network by avoiding voltage band violations or line overloadings, it also becomes crucial to properly observe and foresee voltages and power flows down to the low-voltage level. Hence, this chapter investigates the short-term probabilistic prediction of the state (i.e., power injections, power flows, voltages) of low-voltage grids for operation purposes. To the best of the author's knowledge, this has never been proposed in the literature.

Furthermore, a focus has recently been given to control schemes that explicitly consider uncertainty in their design [366–368]. This appears inevitable to cope with the volatile and hardly predictable nature of loads and DERs in distribution grids. Such design is based on stochastic programming, robust optimization or chance-constrained Optimal Power Flow (OPF) [369–371]. However, proposed approaches generally assume a given error distribution and overlook the large variations in uncertainty over both temporal and spatial dimensions of a system. As presented in Section 9.4, the inherent uncertainty in the state of distribution grids can still be estimated at high temporal and spatial resolutions with reasonable accuracy, even at the LV level. In addition, in contrast to the transmission level, a large majority of control schemes presented in the literature for the distribution grid level rely on corrective measures. This is also the case for voltage control, where the proposed optimal control mechanisms are triggered in reaction to the observation of voltage band violations. The author strongly believes that the preventive estimation of such grid constraint violations together with the design of preventive control measures allow for more cost-efficient distribution grid operation. Consequently, this chapter promotes the use of short-term quantile forecasts for preventive voltage control. Concretely, quantile forecasts of different quantities are directly integrated into a two-stage OPF control scheme. It aims to optimally estimate in advance the required PV power curtailment in order to maintain the near-future voltages within limits and hence avoid more expensive curtailment in real-time.

The remainder of this chapter is structured as follows. Section 9.2 first elaborates on the forecasting workflow followed in this work, especially focusing on the training phase. Next, Section 9.3 discusses the application of deterministic load forecasting algorithms and their evaluation at the level of an LV grid. It analyzes the statistical properties of the deterministic predictions and points out the shortcomings of traditionally used algorithms and evaluation metrics. Subsequently, the notion of short-term probabilistic state forecasting is detailed in Section 9.4. Two different quantile forecasting algorithms are developed, and their performance is evaluated for different levels of PV and EV penetration. Section 9.5 proposes a preventive voltage control scheme that explicitly accounts for uncertainty by integrating quantile forecasts into the corresponding optimization problem. The advantages of quantile forecasts over point forecasts are demonstrated considering different imbalance prices. The main outcomes are finally summarized in Section 9.6 which also introduces various avenues for future work.

9.2 FORECASTING WORKFLOW

In this work, the development of both deterministic and probabilistic forecasting algorithms follows a well-defined workflow which is illustrated in Figure 9.1. The workflow is inspired by common practices among the forecasting community. However, note that the purpose of this chapter is definitely not the development of the best possible forecasting model. In any case, the performance of an algorithm depends on the data set on which it has been tested. As noted by the authors in [128], the notion of universally best technique does not make sense. In fact, this chapter principally focuses on the application of short-term forecasting in distribution grids. Special care is nevertheless given to the design, training, and tuning of the different forecasting algorithms introduced in the following sections. The author is confident that the presented workflow enables a robust design of these algorithms.

First of all, the entire data set is split into a training set and a test set. The training set is used to select the best model of a given forecasting algorithm, which refers to two main aspects: feature selection and hyper-parameter tuning. On the one hand, a forecasting algorithm is rarely fed with the original data but requires a reduced set of specific data called features. The most relevant features are normally selected based on domain knowledge. Multiple combinations of the most relevant features must nevertheless be tested to fine-tune the model. On the other hand, most algorithms contain so-called hyper-parameters which must be defined before the actual learning phase[1]. For example, the number of hidden layers and neurons, the activation function, the learning rate for gradient descent, and the regularization term are typical hyper-parameters of a neural network. Hyper-parameter tuning can be carried out under the form of a grid search. First, multiple possible values are selected for each hyper-parameter. Second, the performance of the algorithm is evaluated for all combinations of hyper-parameter values. Hence, the model selection phase consists of a feature selection step and a

---

[1] At the difference of hyper-parameters, model parameters are defined during the learning phase.

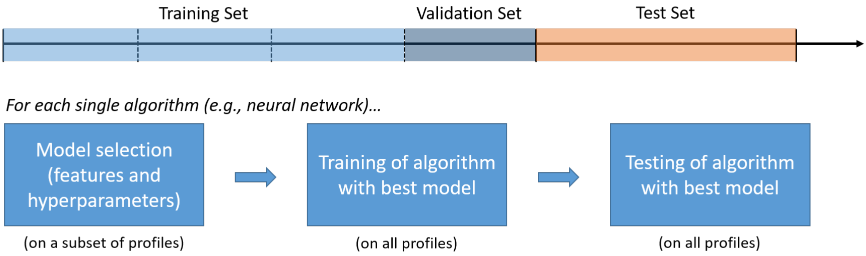

Training Set
Validation Set
Test Set
For each single algorithm (e.g., neural network)...
Model selection (features and hyperparameters)
(on a subset of profiles)
Training of algorithm with best model
(on all profiles)
Testing of algorithm with best model
(on all profiles)

FIGURE 9.1: Workflow followed in this work for the design of a forecasting algorithm.

hyper-parameter tuning step, where multiple combinations must be tested. For that purpose, the original training set is further split into an actual training set and a validation set. The training phase (or learning phase) fits the parameters (e.g., weights) of the model in order to best match the input features with the output (i.e., value to predict). Subsequently, all trained models are evaluated on the validation set, and the model leading to the best accuracy is selected.

Alternatively, model selection can be performed according to the so-called k-fold cross-validation. This procedure mitigates the risks of overfitting, i.e., the selection of a model which is too specific to the training data and fails to generalize to previously unseen data. Concretely, the original training set is split into k subsets, where $k-1$ subsets are used as the actual training set, and the remaining subset is used as the validation set. This process is iterated k times, where each subset is used once as the validation set. Eventually, the best model corresponds to the model associated with the set of features and hyper-parameter values leading to the best average accuracy over the k validation sets.

On a side note, it is commonly known that the selection of features has a stronger influence on the model performance than the choice of hyper-parameter values. Hence, for each cross-validation iteration, the feature selection step is performed in the first stage, using common hyper-parameter values proposed in the literature. Hyper-parameter tuning is performed in the second stage based on the previously selected features to further fine-tune the model. In theory, feature selection and hyper-parameter tuning could be performed simultaneously by adding the multiple combinations of features to the grid search. In practice, this leads to an explosion of the number of grid-search combinations, which is particularly time-consuming and does not bring substantial benefit.

The different algorithms presented in this work are applied to a multitude of profiles. In order to reduce the computational cost, a unique model per algorithm is selected based on a subset of profiles. More precisely, the unique model consists of the set of features and hyper-parameter values leading to the best average accuracy over the subset of profiles. Next, each algorithm is trained on the entire original training set of every single profile. Each trained algorithm is finally evaluated on the test set. The test set enables the comparison of the performance of multiple algorithms. It is important that the data from the test set have not been used in the training process in order not to bias the performance evaluation.

9.3 INADEQUACY OF STANDARD DETERMINISTIC APPROACHES

As a motivation for this section, Figure 9.2 illustrates the 15-minute resolution profile of an aggregation of multiple loads and of a single consumer over one week, together with 24-hour-ahead deterministic forecasts performed by an Adaptive Markov Chain Model (AMCM) and by Support Vector Regression (SVR). More information on the algorithms is given in Section 9.3.1. As already studied in Chapter 5, the single consumer exhibits sharp and hardly predictable consumption peaks. In contrast, the aggregated profile is typical of the measurements at an MV/LV transformer level where the high volatility of individual consumers is partially smoothed out. Visually, both algorithms produce predictions similar to the measurements of the aggregation even though the prediction based on the AMCM is somewhat more volatile. However, the predictions for the single consumer are substantially different. The SVR algorithm produces a very smooth profile which is not characteristic of the volatility observed in the measurements. All consumption peaks are not reflected by the SVR. In terms of point-wise accuracy, it achieves a Mean Absolute Percentage Error (MAPE) of 37%. Conversely, the AMCM-based forecast is a spiky profile. For this particular consumer, the algorithm is able to reproduce a similar load behavior at the weekend but wrongly predicts high activity on Monday afternoon. Although the AMCM seems to give a visually more realistic forecast than the SVR algorithm, it achieves a MAPE of 73%, i.e., twice higher than the SVR accuracy. These observations motivate a deeper investigation of the weaknesses and strengths of various load forecasting algorithms and require a more exhaustive evaluation of their performance.

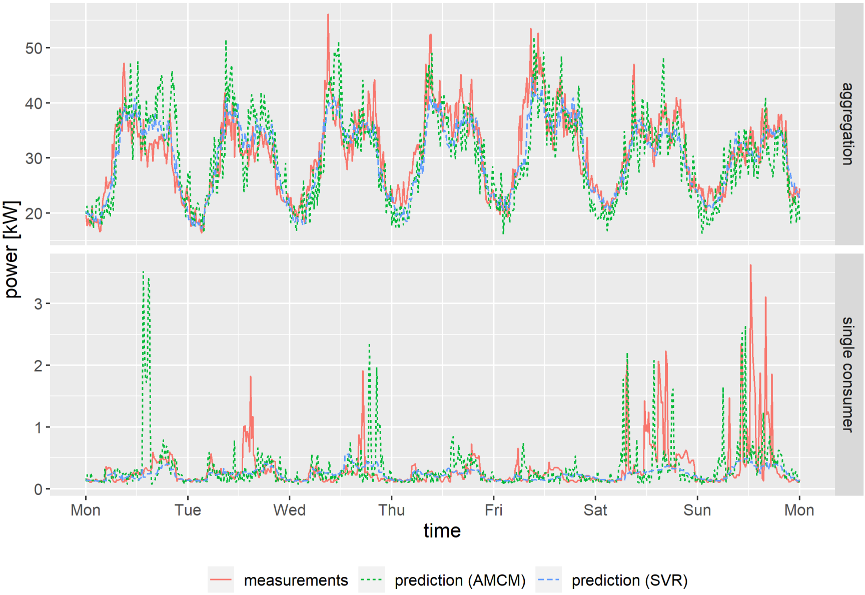


FIGURE 9.2: Measurements and associated predictions based on the AMCM and SVR algorithms for two representative load profiles.

Consequently, four load forecasting algorithms with different characteristics are first presented in Section 9.3.1. Next, Section 9.3.2 introduces different metrics to evaluate the forecast performance. Besides standard point-wise metrics, the ramp score and the adjusted error are alternative metrics for the evaluation of volatile profiles, which are proposed by the authors in [364] and [136], respectively. Additionally, the statistical properties of load forecasts are analyzed. Section 9.3.4 finally evaluates the performance of the four presented forecasting algorithms on a large data set according to the different introduced metrics. The analysis focuses on 24-hour-ahead predictions. More precisely, forecasting is performed at each time step on a rolling basis with a horizon of 24 hours[2] [10].

---

[2] Note that day-ahead forecasting would be the most intuitive approach under the conditions that SM measurement data are traditionally sent once a day to the main utility server. In other words, day-ahead forecasts are issued at a certain point in the day for all time steps of the next day. Nevertheless, such an approach might lead to a decreasing performance over the day due to the increasing prediction horizon, which is not desired in this work.

9.3.1 *Forecasting Algorithms*

For the purpose of this work, the four following forecasting algorithms are considered: Persistence method, Auto-Regressive Moving Average model with eXogenous inputs (ARMAX), AMCM, and SVR. Except for the AMCM algorithm, these different approaches are extensively used in the literature on time series forecasting. This set of forecasters can obviously be extended to a much larger variety of additional algorithms, notably more advanced approaches. The authors in [4] and [9] review the algorithms generally used in the load forecasting literature and for the short-term prediction of building energy consumption, respectively. In addition, the authors in [372] comprehensively presents and compare seven popular load forecasting algorithms applied to the building level and point out the difficulty of predicting individual residential loads. Nevertheless, this study does not focus on the forecasting algorithms themselves but on the characteristics of their predictions, which can already be properly covered by the four selected algorithms. On the one hand, the ARMAX and SVR algorithms are designed to minimize the point-wise error between the measurements and the forecasts at each time step and output relatively smooth profiles. Although they are structurally very different, they both achieve good accuracy on smooth measurement profiles but perform poorly on volatile loads such as single households [123]. Similar characteristics can be observed with linear regression models, gradient boosting trees, neural networks, and deep neural networks, although the former appear to better grasp the inherent uncertainty of load profiles at the end-user level [373, 374]. On the other hand, the persistence model and the AMCM intend to keep the original statistical properties, but at the detriment of the point-wise accuracy. The most suitable features and hyper-parameters for the ARMAX, AMCM, and SVR algorithms have been selected by grid search with k-fold cross-validation on a training set covering up to 2 years of data.

9.3.1.1 *Persistence Method*

The persistence model is a simple method that does not require any learning. It is often used as a benchmark and assumes that the value to forecast is equal to the last observable value:

$$\hat{y}_t = y_{t-H}, \tag{9.1}$$

where $\hat{y}_t$ and y_{t-H} are the predicted and true values at time t and $t-H$, respectively. In addition, H corresponds to the forecasting horizon, which is equal to 24 hours (i.e., 96 time steps with 15-minute resolution) in the context of this study.

9.3.1.2 *Auto-Regressive Moving Average Model with Exogenous Inputs*

The ARMAX model decomposes a time series into an Auto-Regressive (AR) and a Moving Average (MA) part, and can integrate exogenous variables. It is widely used by utility providers because of their relatively fast and easy implementation with suitable accuracy on highly aggregated data. The ARMAX model used in this study is given by:

$$\begin{aligned} y_t = &\sum_{i=1}^{p} \psi_i \cdot y_{t-i} + \sum_{j=1}^{q} \theta_j \cdot \epsilon_{t-j} \\ &+ \sum_{n=1}^{N} \left[\alpha \cdot \sin\left(\frac{2\pi \cdot n \cdot t}{m}\right) + \beta \cdot \cos\left(\frac{2\pi \cdot n \cdot t}{m}\right)\right] + \epsilon_t, \end{aligned} \tag{9.2}$$

where y_t is the value of time series y at time t, and ϵ_t is the white noise error at time step t. In addition, p and q are the orders of the AR and MA components, and ψ_i and θ_j are the parameters of the AR and MA components, respectively. Finally, N is the order of Fourier terms, α and β are the parameters of the Fourier decomposition, and m is the number of time steps per period. For smart meter data, it is reasonable to fix the period to one day, which gives $m := 96$ with a temporal granularity of 15 minutes. After grid search, the remaining hyper-parameters are set to $p := 1$, $q := 1$, and $N := 8$. The model is trained on the most recent observations spanning over one week.

ARMAX models can be extended with an integral component when a certain trend is observed in the time series. This is however not the case for smart meter data in the short time frame of a few days. Besides, the standard Auto-Regressive Moving Average (ARMA) model fails to capture the long seasonal periods in smart meter data, even with considerably large AR and MA orders. Hence, the authors in [375] recommend decomposing the time series into its first Fourier terms and integrate them as exogenous inputs in the model.

9.3.1.3 *Adaptive Markov Chain Model*

The AMCM is an algorithm based on Markov chains and presented in Section 6.2.2.3, suggested for the synthesis of realistic load profiles at the end-user level. As a reminder, the elements of a traditional Markov transition matrix are real numbers indicating the probability to go from one state at time t to another state at time $t+1$. The core concept of the AMCM is to substitute each transition value with a logistic regression function which accounts for calendar features in the calculation of transition probabilities. Hence, the algorithm is designed to reproduce the statistical properties of the training data, especially the periodicity and probability distribution of consumption values. For the sake of this study, the original algorithm is adapted for prediction purposes by adding weather and historical values to the set of features. Hence, each element of the adaptive transition matrix is given by:

$$h_\theta\left(x_t\right) = g\left(\theta_t^{\mathrm{T}} x_t\right), \quad \text{with } g\left(z\right) = \frac{1}{1+e^{-z}}, \tag{9.3}$$

where x_t is a vector of calendar features, weather data, and historical observations with respect to time t. More precisely, calendar features are one-hot encoded categorical variables that correspond to the hour, the weekday, and the month at time t. Weather data are the temperature and irradiation values at time t. Historical observations refer to the observed consumption values at times $t-i,\ i \in \{H, H+1, H+2, W, W+1, W+2\}$, where H and W are the number of time steps in 1 day and 1 week, respectively. The choice of all features results from the feature selection process in the training phase. Finally, θ_t is a vector of coefficients defined by the training process. The algorithm is trained over a period of two years.

9.3.1.4 *Support Vector Regression*

The SVR model is the regression version of the popular Support Vector Machine (SVM). The SVM algorithm is commonly used for supervised classification problems. It aims to separate input data based on certain characteristics using a virtual boundary, also called hyper-plane, as a delimiter between two classes. The objective of the SVM algorithm is to draw this hyper-plane with the widest possible margin. The data points which are the closest to the hyper-plane and consequently influence its position and orientation are called support vectors. Instead of separating the input data, the SVR version intends to pass the most precise hyper-plane through the data, converting the support vector margin into the smallest possible error margin, ideally containing all the data points. Formally, this consists of solving a convex optimization problem which is commonly written as:

$$
\begin{aligned}
\min_{w,\xi,\xi^*} \quad & \frac{1}{2} \cdot w^\intercal w + C \sum_{t=1}^{T} (\xi_t + \xi_t^*), && (9.4a) \\
\text{s.t.} \quad & w^\intercal \phi(x_t) - y_t \geqslant \varepsilon + \xi_t, && (9.4b) \\
& y_t - w^\intercal \phi(x_t) \geqslant \varepsilon + \xi_t^*, && (9.4c) \\
& \xi_t, \xi_t^*, \varepsilon \geqslant 0, \quad C > 0, \quad \forall t \in \{1, ..., T\}, && (9.4d) \\
& k(x_i, x_j) = \phi(x_i)^\intercal \phi(x_j), && (9.4e)
\end{aligned}
$$

where x_t is a vector of input features at time t and y_t is the predicted value at time t. In addition, w is a vector of weights related to the input features, C is a parameter that balances the importance of both terms in the objective function, ε defines the margin of insensitive loss, and T is the total number of time steps in the training set. Next, ξ_t and ξ_t^* are slack variables introduced to ensure feasibility. Finally, $k(x_i, x_j)$ is a chosen kernel function and $\phi(x_i)$ is the corresponding feature map. Kernel functions are also described as efficient inner products. With the help of the feature map, the features are mapped into an inner product space induced by the kernel function. In this new space, the computation occurs in a more efficient manner. The goal is to find a function that stays within a ε-margin from the predicted values y, but also exhibits flatness. The first requirement is met through minimization of the so-called ε-insensitive loss. The second requirement is reflected in the l_2-regularization of the weights. There is an extensive literature on the SVR algorithm and more information can be found in [376–378].

The SVR algorithm has proven to perform particularly well in the short-term load forecasting literature [9, 362]. For the purpose of this study, a training period of 1 year appears to be sufficient. Similar features as for the AMCM-based approach are used. The SVR model comes from the *scikit-learn* python library and is based on the Radial Basis Function (RBF) kernel. After hyper-parameter tuning, C and ε are set to 1000 and 1, respectively.

9.3.2 *Deterministic Evaluation Metrics*

In the load forecasting literature, algorithms are generally assessed via point-wise metrics which compare the true and predicted value at each time step [9]. This is the case of standard metrics such as the Mean Absolute Percentage Error (MAPE) and the Normalized Root Mean Square Error (NRMSE). Alternatively, the authors in [136] suggest the use of an adjusted error metric

load type	algorithm	MAPE	NRMSE	adjusted NRMSE (2h | 4h)	ramp score
aggregation	AMCM	13%	17%	11% | 10%	4.5%
aggregation	SVR	8.2%	12%	9.2% | 8.8%	1.9%
single consumer	AMCM	73%	147%	102% | 93%	3.4%
single consumer	SVR	37%	113%	108% | 106%	1.9%

Table 9.1: Performance evaluation of the forecasts of the two representative load profiles illustrated in Figure 9.2.

that mitigates the double penalty effect induced by standard point-wise metrics. Besides, the ramp score is suggested by the authors in [364] to assess the prediction accuracy of ramp events, which is of particular interest in solar forecasting. These various metrics are defined in the following. Their application to the different predictions of Figure 9.2 are presented in Table 9.1. Note that this section proposes a non-exhaustive list of metrics for comparing two time series, but other metrics (e.g., based on Dynamic Time Warping (DTW) [379]) might also be suitable.

9.3.2.1 *Standard Point-Wise Metrics*

The MAPE and the NRMSE are among the most common metrics for the evaluation of deterministic forecasts. They are defined as:

$$\text{MAPE} = \frac{100\%}{T} \cdot \sum_{t=1}^{T} |\frac{y_t - \hat{y}_t}{y_t}|, \tag{9.5a}$$

$$\text{NRMSE} = \frac{100\%}{\bar{y}} \cdot \sqrt{\frac{1}{T} \cdot \sum_{t=1}^{T} (y_t - \hat{y}_t)^2} = \frac{100\%}{\bar{y}} \cdot \|y - \hat{y}\|_2, \tag{9.5b}$$

where y_t and $\hat{y}_t$ are the true and predicted values at time t, respectively. In addition, $\bar{y}$ is the mean value over all observations, and T is the number of time steps. The lower the error metric value, the better the forecast accuracy. Since they are given in percent, the MAPE and NRMSE allow for comparison of the forecast accuracy not only between different algorithms on a same profile, but also between different profiles. By definition, large deviations are penalized to a greater extent by the NRMSE than by the MAPE.

As shown in Table 9.1, standard point-wise metrics penalize the AMCM-based approach which creates profiles with high volatility and seem to favor the smooth predictions of the SVR algorithm. This is the case for the illustrated small consumer. Although they might be of realistic magnitude, the power spikes created by the AMCM-based approach do not exactly match in time with the observed spikes in the actual load profile. This leads to the so-called "double penalty effect" in the performance evaluation via commonly used metrics. In fact, these point-wise metrics highly penalize the volatile forecasts both at the time steps when the true spike is not predicted and at the time steps with wrongly predicted spikes.

9.3.2.2 *Adjusted Error Metric*

In order to mitigate the double penalty effect, the authors in [136] propose an adjusted error measure that allows for small, possibly discontinuous, displacements of the predicted values in time. The principle of the adjusted error metric has already been described in Section 7.3.3.2 for the Root Mean Square Error (RMSE). For comparison purposes between different load profiles, this work makes use of the adjusted NRMSE which is defined as:

$$\text{NRMSE}^{\omega} = \min_{P \in \mathcal{P}} \quad \frac{100\%}{\bar{y}} \cdot \|y - P \cdot \hat{y}\|_2, \tag{9.6a}$$

$$\text{s.t.} \quad P_{uv} = 0, \quad \forall\ |u - v| > \omega, \tag{9.6b}$$

where $\omega \geqslant 0$ is an adjustment limit, P is a permutation matrix, $P_{uv} \in P$ refers to the displacement of the estimated value $\hat{y}_u$ from time step u to time step v, and $\mathcal{P}$ is the complete set of restricted permutations. In this sense, a good load forecasting algorithm is able to accurately predict consumption values within a time window of $\pm\omega$ time steps. According to the adjusted NRMSE, Table 9.1 shows that the AMCM-based approach outperforms the SVR algorithm for the small consumer in question, which contrasts with the evaluation via the standard point-wise metrics. The relative difference in accuracy is even stronger when ω is set to 4 hours instead of 2 hours. At the aggregate level, even if the difference in performance is reduced, the SVR forecast is still more accurate.

9.3.2.3 *Ramp Score*

The prediction of sudden and significant changes in load might be of particular interest. For that purpose, the authors in [364] propose the ramp score, which evaluates the ability to predict significant ramp events. This score

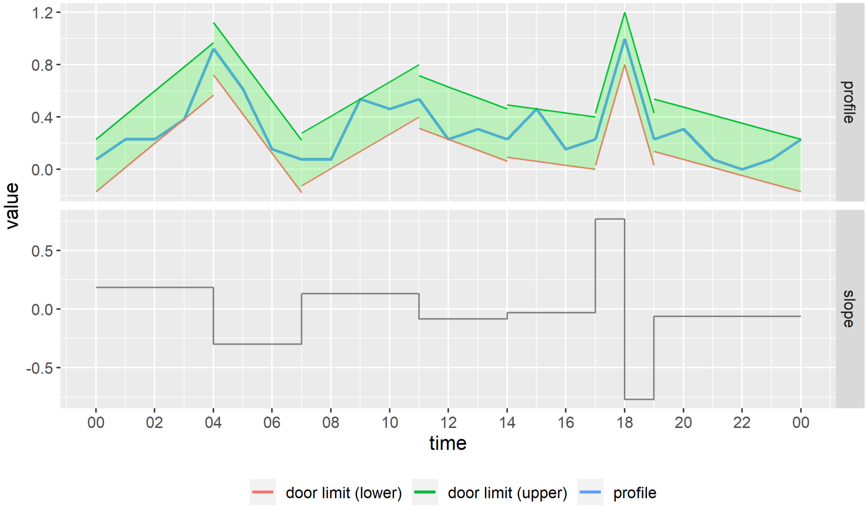


FIGURE 9.3: Principle of the swinging door algorithm with $\epsilon := 0.4$ and resulting extracted ramps.

is principally used in the solar forecasting literature. It is based on the swinging door algorithm suggested in [380] for data compression purposes. As illustrated in Figure 9.3, this algorithm performs a piece-wise linear approximation of a profile over multiple time steps and requires a sensitivity parameter ϵ that controls the significance of the resulting ramps (i.e., width of the door). As preparation for the ramp score, the swinging door algorithm is applied to the true and predicted profiles, and the slope magnitude of the resulting piecewise linear approximations is extracted at each time step. Consequently, the ramp score is defined as:

$$\text{ramp score} = \frac{1}{T} \cdot \sum_{t=1}^{T} |\text{slope}(y_t) - \text{slope}(\hat{y}_t)|, \tag{9.7}$$

where y_t and $\hat{y}_t$ are the true and predicted values at time t, respectively. In addition, 'slope' refers to the slope of the ramp defined by the swinging door algorithm, and T is the number of time steps. The lower the score, the better the ramp forecast accuracy. For comparison purposes, the true profiles are scaled by min-max normalization between zero and one, and the same scaling is applied to predicted values. In Table 9.1, ϵ is set to 0.1, which approximately leads to ten times fewer detected ramps than the number of

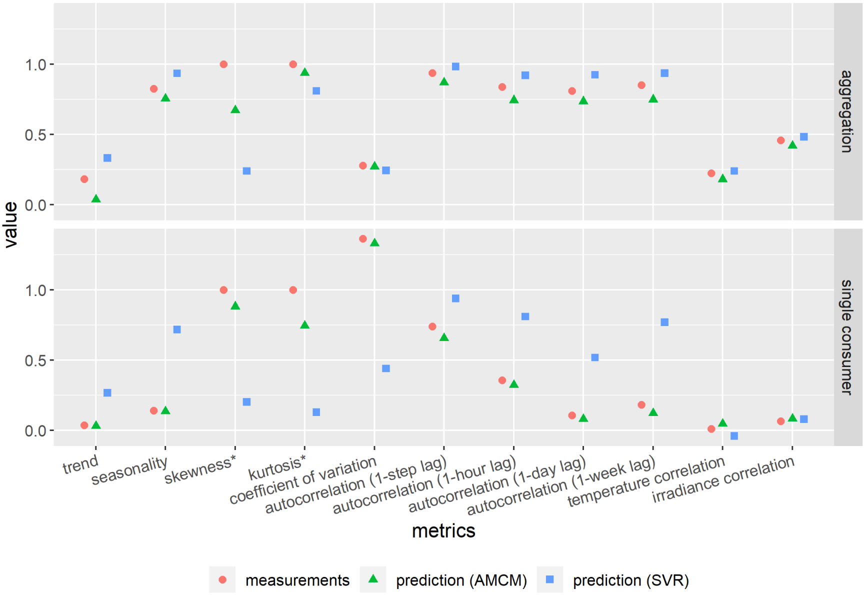


FIGURE 9.4: Various statistical properties of the load profiles and respective predictions presented in Figure 9.2. The skewness and kurtosis values are scaled by the respective values calculated over the measurements.

time steps. It appears that the ramp score is in line with standard metrics, notably for the single consumer.

9.3.3 *Statistical Properties*

In addition to the consideration of various evaluation metrics, this work also studies the statistical properties of the forecasts in comparison with the measured time series. In the following, the trend and the seasonality, the skewness and the kurtosis, the Coefficient of Variation (CV), the autocorrelation, and the correlation with weather variables are introduced. Figure 9.4 visualizes the statistical properties of the AMCM and SVR predictions visible in Figure 9.2, in comparison with the measured load profiles.

9.3.3.1 *Trend and Seasonality*

A time series can be decomposed into three additive components as follows:

$$y = T_y + S_y + I_y, \tag{9.8}$$

where y is the original time series, and T_y, S_y, and I_y refer to the trend, seasonal, and irregular components, respectively. The trend is defined as the centered moving average with a fixed time window which is set to one week in this study. The seasonality is calculated as the average profile over a fixed time period which is also set to one week. The irregular component is commonly referred to as noise. In order to quantify the level of trend and seasonality in a given time series, the authors in [381] suggest following measures:

$$\text{trend} = 1 - \frac{\text{var}(I_y)}{\text{var}(y - T_y)}, \tag{9.9a}$$

$$\text{seasonality} = 1 - \frac{\text{var}(I_y)}{\text{var}(y - S_y)}, \tag{9.9b}$$

where ‘var’ refers to the variance of the corresponding time series. As illustrated in Figure 9.4, the original profiles of the two representative loads barely exhibit a trend. This is partially captured by both forecasting algorithms, although the SVR algorithm slightly overestimates the trend. Furthermore, they both properly reflect the high level of seasonality observed in the aggregated load, but the SVR algorithm largely overestimates the seasonal effect in the behavior of the single consumer.

9.3.3.2 *Skewness and Kurtosis*

The skewness and kurtosis describe the shape of the probability distribution of a time series. While the skewness measures the asymmetry around the mean of the probability distribution, the kurtosis refers to the sharpness of its peak. They are given by:

$$\text{skewness} = \frac{1}{T \cdot \sigma^3} \cdot \sum_{t=1}^{T} (y_t - \bar{y})^3, \tag{9.10a}$$

$$\text{kurtosis} = \frac{1}{T \cdot \sigma^4} \cdot \sum_{t=1}^{T} (y_t - \bar{y})^4, \tag{9.10b}$$

where σ is the standard deviation of time series y[3]. A distribution skewed to the left has a negative value of skewness, and vice versa. A normal (or Gaussian) distribution has a skewness value of zero and a kurtosis value of three, whereas a higher kurtosis reflects a sharp distribution peak. Smart meter data are normally positively skewed and exhibit a high kurtosis, which is representative of values that are principally concentrated close to the mean with a few high spikes. In order to stay within the range of the other metrics under consideration, the skewness and kurtosis values displayed in Figure 9.4 are scaled by the respective values calculated for the measured load profiles. Both algorithms underestimate the skewness of the two load profiles, which is even more pronounced for the SVR algorithm. The latter also clearly fails to represent the typically high kurtosis of the single consumer.

9.3.3.3 *Coefficient of Variation*

The Coefficient of Variation (CV) is defined as the ratio between the mean and the standard deviation and can be interpreted as the volatility of a profile. This quantity is also shown in Figure 9.4. While the low volatility of the aggregated load is perfectly recognized by both algorithms, the high volatility of the single consumer is largely underestimated by the SVR algorithm.

9.3.3.4 *Autocorrelation*

The autocorrelation represents the Pearson's correlation of a time series with a delayed copy of this same time series. Figure 9.4 illustrates this value for the different load profiles under consideration with a lag of one time step, one hour, one day, and one week. The high autocorrelation of the aggregated measured profile is well respected by the prediction algorithms. However, unlike the AMCM, the SVR algorithm overestimates this quantity for the single consumer. The mismatch even rises with increasing lag.

9.3.3.5 *Correlation Coefficient*

Finally, the Pearson's correlation coefficient with the profiles of temperature and solar irradiance is considered and illustrated in Figure 9.4. The correlation with weather data is barely perceptible for the single consumer, which increases at the aggregation level. In any case, these properties are well captured by both algorithms.

---

[3] Remaining variables and parameters are defined in the preceding equations.

9.3.4 *Performance Evaluation*

The preliminary analysis on the representative load profiles from Figure 9.2 indicates that the SVR algorithm outperforms the AMCM-based approach at the aggregate level with respect to all evaluation metrics under consideration. Both studied algorithms also properly reflect a large majority of statistical properties for the aggregated load. Nevertheless, mixed results appear for load forecasting at the end-user level. While the standard point-wise metrics and the ramp score still suggest better performance for the SVR algorithm, the adjusted error metric points out its limits and recognizes the more realistic forecasting outcome of the AMCM-based approach. In terms of statistical properties, the AMCM-based approach definitely depicts a more faithful representation of reality. In fact, each metric characterizes a different facet of the forecasts, and, according to the metric, the respective performance of different prediction algorithms can substantially vary.

9.3.4.1 *Case Study*

In order to draw more solid conclusions, the analysis is extended to a statistically significant data set of 1’000 load profiles, where the four algorithms presented in Section 9.3.1 are evaluated. Concretely, the data set used for training and testing the algorithms consists of residential and commercial load profiles with a temporal resolution of 15 minutes coming from the distribution grid of the City of Basel. The data set is described in Section 3.2.1 and its preparation is detailed in Section 4.3. Out of a total of 20’000 smart meter profiles, 1000 aggregations of 1 to 30 randomly chosen individual consumers are created. This typically represents the load visible at the nodes of low-voltage grids, which is characterized by high volatility and low predictability. Moreover, weather data are provided by a meteorological station of MeteoSwiss in the City of Basel [237]. Due to the absence of weather forecasts, it is important to notice that the actual weather quantities at the time of prediction are used in this study. This assumption of a perfect weather forecast is definitely wrong in reality. Nevertheless, it is not expected to significantly affect the outcome of this study, especially since all algorithms have access to the same information in theory.

9.3.4.2 *Results*

All algorithms are evaluated over a period of one full year, obviously different from the training and validation periods. Figure 9.5 first illustrates the

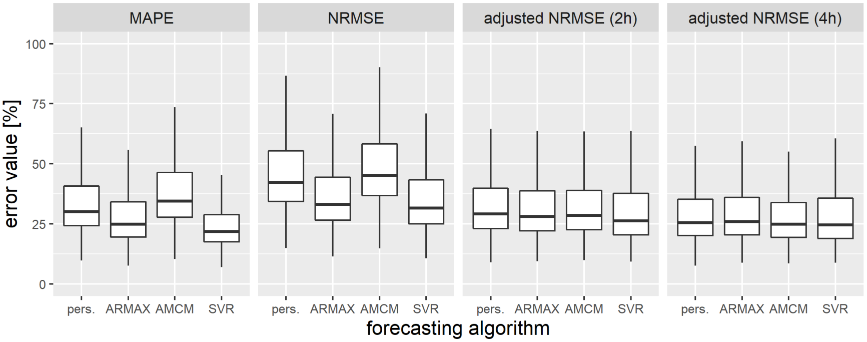


FIGURE 9.5: Performance evaluation of different forecasting algorithms based on the MAPE, NRMSE, and two variants of the adjusted NRMSE.

accuracy of the prediction algorithms with respect to the standard MAPE and NRMSE, as well as the adjusted NRMSE with an adjustment limit set to 2 hours and 4 hours. Better point-wise accuracy is confirmed by both standard metrics for the ARMAX model and especially the SVR algorithm, which are designed for that purpose. Nevertheless, standard metrics give a biased image of the load forecasting accuracy at low aggregation levels, where the double penalty effect can be suspected. By allowing slight time displacements of the predicted values before evaluation, the adjusted NRMSE shows that all algorithms finally exhibit very similar accuracy on the large data set. The SVR algorithm is still slightly more effective when the adjustment limit is set to 2 hours, which stays in line with the standard NRMSE but leads to the same accuracy as the competing algorithms with a 4-hour adjustment limit. In fact, all algorithms are not capable to correctly approximate the peaks in consumption within a suitable time window. This confirms the difficulty of any standard point forecasting algorithm to predict volatile loads and reveals their limits.

In Figure 9.6, the ramp score is shown for ϵ values of 0.05 and 0.1. A lower ϵ value leads to a higher number of detected ramps and a generally higher score. It appears as if the ARMAX and SVR models better approximate significant ramps, but this is again due to the double penalty effect which is transmitted to the ramp profiles. In fact, the ramp score is still a point-wise comparison of significant changes in load whose exact time is hardly predictable by any 24-hour-ahead forecasting algorithm. Due to the high load volatility, the ramp score also gives a biased image of the forecasting accuracy.

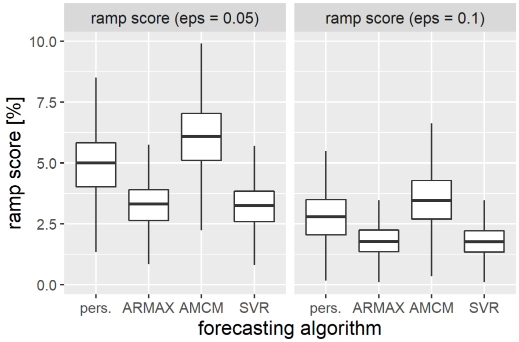


FIGURE 9.6: Performance evaluation of different forecasting algorithms based on two variants of the ramp score.

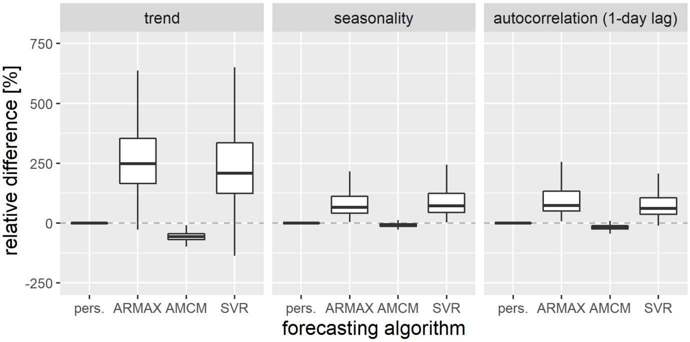


FIGURE 9.7: Performance evaluation of different forecasting algorithms based on the relative difference in the trend, seasonality, and autocorrelation with a lag of 1 day.

Finally, the forecasting algorithms are evaluated from the perspective of statistical properties. The respective metrics are defined as the relative difference between the statistical value of the forecast and the statistical value of the measurement:

$$\text{metric}_{\text{stat}} = 100\% \cdot \frac{\hat{\chi} - \chi}{\chi}, \tag{9.11}$$

where χ and $\hat{\chi}$ are the statistical values of the measurement and the forecast, respectively. On this basis, Figure 9.7 first considers the trend, seasonality, and autocorrelation with a lag of 1 day. By definition, the persistence method

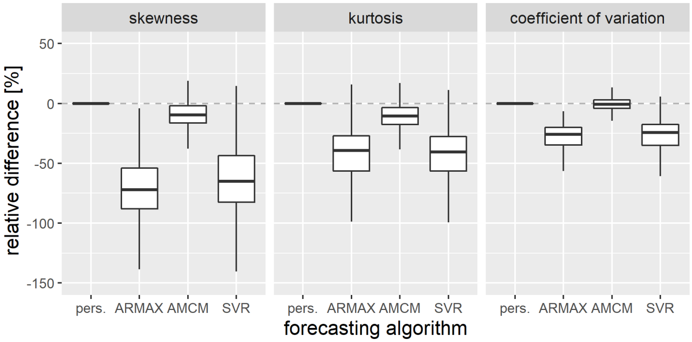


FIGURE 9.8: Performance evaluation of different forecasting algorithms based on the relative difference in the skewness, kurtosis, and coefficient of variation.

perfectly grasps all statistical properties. The seasonality and autocorrelation are also particularly well preserved by the AMCM-based approach, which seems to somewhat underestimate the trend in terms of relative difference. In contrast, all three properties are overestimated by the ARMAX and the SVR algorithm. It must nevertheless be noted that the trend, seasonality, and autocorrelation values for the measurements are especially low such that the relative difference metric should be interpreted with caution. Besides, the relative difference in skewness, kurtosis, and coefficient of variation is displayed in Figure 9.8. While these properties are accurately reflected by the AMCM-based forecasts, they are substantially underestimated by the ARMAX and the SVR algorithm.

9.3.4.3 *Discussion*

To sum up, the forecasting literature almost exclusively relies on standard metrics such as the MAPE or RMSE for the evaluation of deterministic load forecasts. However, these metrics only focus on one specific aspect, the point-wise error, which can be problematic at the level of end-consumers with highly volatile load profiles. In fact, point-wise error metrics favor very smooth and unrealistic predictions while doubly penalizing volatile forecasts. Such double penalty effect could be mitigated with the help of the adjusted error metric presented in [136]. In addition, this analysis highlights a considerable gap between the outcome of traditional prediction algorithms and more volatile forecasts. On the one hand, traditional prediction algorithms purely seek point-wise accuracy at the detriment of statistical properties. On the other hand, volatile forecasts exhibit good statistical properties but have a limited consideration of the time dimension. In any case, aside from the notion of pure point-wise accuracy, advanced deterministic algorithms cannot outperform the very simple persistence method. This also holds true for the proposed AMCM algorithm that is designed to learn the time periods with statistically higher activity. In fact, the behavior of residential and commercial end-users is too uncertain to give a unique, accurate estimate of their load at each time step.

In this context, the actual usefulness of deterministic load forecasting at the level of end-users or low-voltage nodes is questionable. It must be noted that the proposed analysis focuses on 24-hour-ahead forecasting. It is reasonable to expect that the forecasting accuracy shall increase with decreasing forecasting horizon. For example, the authors in [124] observe a slight drop of the NRMSE at the level of single consumers when the forecasting horizon is reduced to one hour and, to a lesser extent, two hours. At horizons larger than two hours, the most recent observations are however not especially valuable anymore such that the prediction error remains relatively constant, independently from the horizon. This phenomenon is also observed by the authors in [358]. In any case, the load uncertainty at this level is so high that point forecasts can only provide incomplete information. Hence, a comprehensive load forecasting approach at the level of low-voltage grid nodes must inevitably quantify this uncertainty. This is the case of probabilistic forecasting algorithms, as presented and exploited in the next section.

9.4 SHORT-TERM PROBABILISTIC STATE FORECASTING

The main idea of this section is the development of a so-called short-term probabilistic state forecaster that estimates the near-future state uncertainty in LV grids.h In fact, probabilistic forecasting algorithms enable a comprehensive prediction by covering the entire uncertainty range. They generally do not assume any error distribution and can update their prediction in real-time [128]. For example, the authors in [137, 382] present novel regression neural network and Long Short-Term Memory (LSTM) algorithms for quantile load forecasting with promising results. However, the literature mainly focuses on load forecasting and largely disregards the direct prediction of other quantities such as the voltage and line loading. As repeatedly mentioned over the course of this thesis, the knowledge of the system state at the LV level is all the more relevant for a cost-efficient operation of distribution grids. To the best of the author's knowledge, the probabilistic forecasting of voltages and power flows at the level of an LV grid has never been presented in the literature.

Consequently, a quantile neural network and a novel probabilistic version of the k-Nearest Neighbor algorithm are designed in this section to predict the net power consumptions, power flows, and bus voltage magnitudes at the LV level. The forecasting algorithms are described in Section 9.4.1. As presented in Section 9.4.2.1, the case study relies on an LV section of the distribution grid of the City of Basel and considers multiple levels of PV and EV penetration. The probabilistic forecasting performance is evaluated in Section 9.4.2. This analysis also quantifies the added value of real-time instead of time-delayed SM measurements and of an additional feature that indicates at which points in time the EVs are charging.

9.4.1 *Methodology*

This section details the methodology behind probabilistic forecasting. For the purpose of this work, probabilistic forecasting is represented by quantile forecasts. In the following, a quantile Neural Network (NN) and a quantile K-Nearest Neighbor (KNN) are defined, after briefly presenting the concept and setup of quantile forecasting.

9.4.1.1 *Concept and Setup of Quantile Forecasting*

In deterministic forecasting (or point forecasting), the algorithm outputs only one value which is the most probable value on the basis of the input features. More concretely, the predicted value is the value that shall lead to the lowest error in comparison with the true value. In fact, predictions are never perfect but are associated with a certain uncertainty, which can be represented by a probability distribution of the possible forecast values. In this context, quantile forecasting refers to the prediction of a given quantity at various quantiles of this probability distribution. The training process consists of minimizing the pinball loss function, which creates separate forecasts for the different quantiles:

$$J_q = \frac{1}{N} \sum_{n=1}^{N} \begin{cases} (y_n - \hat{y}_{q,n})\, q & \text{if } y_n \geqslant \hat{y}_{q,n}, \\ (\hat{y}_{q,n} - y_n)\,(1-q) & \text{if } y_n < \hat{y}_{q,n}, \end{cases} \tag{9.12}$$

where J_q is the pinball loss function for quantile q, y_n is the target value of sample n, $\hat{y}_{q,n}$ is the predicted value for quantile q of sample n, and N is the number of samples in the training set. The loss function is asymmetric such that for any quantile higher than the 50%-quantile, forecasting errors due to underestimation of the target value get penalized more than the errors due to overestimation, and vice-versa. Note that the 50%-quantile corresponds to the usual point forecast.

The quantile forecasting workflow is carried out as described in Section 9.2. In the same system, different quantities and quantiles are correlated and could be predicted simultaneously. However, a preliminary study suggests that the use of one single ML model per grid component, per quantity, and per quantile is more efficient in terms of computation time and accuracy, as long as the most influencing quantities belong to the feature set.

9.4.1.2 *Quantile Neural Network*

An Artificial Neural Network (ANN), also referred to as Neural Network (NN), is a mathematical model that is designed to approximate any non-linear function. Its working principle is inspired by the structure of the human brain. The model is basically a weighted linear combination of neurons organized in layers, where each neuron is a non-linear function. A classical neural network comprises an input layer, one or multiple hidden layers, and an output layer. The number of neurons in the input layer is given by the number of features. The neurons of the input layer are connected to the neurons of the first hidden layer. The number of hidden layers and the number of neurons within these hidden layers are hyper-parameters of the model and are set during the model selection phase. The neurons of the last hidden layer are connected to the neurons of the output layer. In a regression problem, the number of neurons in the output layer is determined by the number of target values to be predicted. In this case, the proposed model only consists of one output neuron that refers to a given grid component, a given quantity, and a given quantile. Formally, hidden layers are defined as follows:

$$a^{(1)} = g(\Theta^{(1)} \cdot \begin{bmatrix} 1 \\ x \end{bmatrix}), \tag{9.13a}$$

$$a^{(j)} = g(\Theta^{(j)} \cdot \begin{bmatrix} 1 \\ a^{(j-1)} \end{bmatrix}), \tag{9.13b}$$

where $g(.)$ is an activation function, $a^{(j)}$ is a vector of activation functions in layer j, $\Theta^{(j)}$ is a matrix of weights for the transition between layer $j-1$ and layer j, and x is a vector of input features. Note that each layer contains a bias term of value 1. There are different possibilities for the choice of the activation function. The Rectified Linear Unit (RELU) function and the Exponential Linear Unit (ELU) function count as the most popular activation functions in the literature and are given as follows:

$$\mathrm{RELU}(z) = \max(0, z), \tag{9.14a}$$

$$\mathrm{ELU}(z) = \begin{cases} z & \text{if } z \geqslant 0, \\ e^z - 1 & \text{if } z < 0, \end{cases} \tag{9.14b}$$

where z is a real number. The ELU function is a modification of the RELU function which uses the exponential function to process the input value instead of setting it to zero for negative values. During the training process, the weights of the neural network are adjusted in order to minimize the pinball loss function defined in Equation (9.12). Normally, the cost function also includes a regularization term which prevents the algorithm from overfitting. The factor associated with the regularization term is an hyper-parameter to be tuned. The drop-out of a certain share of randomly chosen neurons during the training process is another way to avoid overfitting. There is extensive literature on ANNs, notably regarding different variants and their training via backpropagation. This is however out of the scope of this thesis. More information can be found in [383, 384].

Depending on the outcome of the hyper-parameter tuning, the ANN models have following structure: 4 hidden layers with 200 neurons per layer, or 6 hidden layers with trapezoidal shape (i.e., n_f, $2n_f$, $3n_f$, $2n_f$, n_f, and $n_f/2$ neurons per respective layer, with n_f indicating the number of input features). Both ELU or RELU activation functions are taken into account during the hyper-parameter tuning phase. In addition, l_2-regularization and a drop-out rate between 0% and 10% are considered to prevent overfitting.

9.4.1.3 *Quantile K-Nearest Neighbor*

Figure 9.9 illustrates the underlying idea of the KNN algorithm used for quantile forecasting, assuming a forecasting horizon of one time step. As starting point, the current state defined by a set of features can be observed. Consequently, the algorithm assumes that if a similar grid state has been observed at a certain point in the past, the current target value of a given quantity is similar to the target value after the similar state found in the past. In this work, the similarity measure is based on the weighted Minkowski distance:

$$d_{t,m} = \|(w^{\text{mink}})^\intercal \, (x_m - x_t)\|_p, \quad \forall m \in \mathcal{M}_t, \tag{9.15}$$

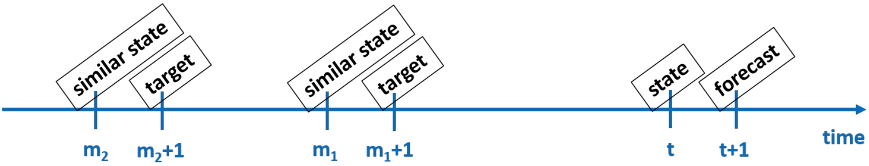


FIGURE 9.9: Principle of the KNN algorithm used for quantile forecasting.

where x_m and x_t are the feature vectors of past sample at time m and of current sample (i.e., at time t), respectively. In addition, w^{mink} is a vector of weights, $\mathcal{M}_t$ is the set of past samples according to current time t, and $p \in \{1, 2\}$ is a hyper-parameter defining the norm. Hence, $d_{t,m}$ is the Minkowski distance between the past system state defined by x_m and the current system state defined by x_t. The Minkowski weights are set according to the importance of the corresponding features, which is given by Lasso cross-validation in this work. After calculating all Minkowski distances (i.e., $\forall m \in \mathcal{M}_t$), the corresponding target values of the k states showing the smallest distances (i.e., k nearest neighbors) are linearly combined to generate a point forecast:

$$\hat{y}_t = \sum_{i=1}^{k} w_i^{\text{knn}} \cdot y_i^{\text{knn}} \tag{9.16}$$

where $\hat{y}_t$ and y_i^{knn} are the predicted value issued at time t and the target value associated to state i, respectively. In addition, w_i^{knn} is the weight associated to state i, and $k \leqslant |\mathcal{M}_t|$ is the number of nearest neighbors. The k nearest neighbors are sorted from the closest neighbor to the furthest neighbor.

At this stage, a suitable number of neighbors as well as optimal KNN weights must still be defined. For that purpose, ten small optimization problems for $k \in \{5, 10, \dots, 50\}$ are solved over the training set:

$$J_{\text{knn}} = \min_{w^{\text{knn}}} \quad \frac{1}{N} \sum_{n=1}^{N} |(w^{\text{knn}})^\intercal y_n^{\text{knn}} - y_n|, \tag{9.17a}$$

$$\text{s.t.} \quad w_i^{\text{knn}} \geqslant 0, \quad \forall i \in \{0, \dots, k\}, \tag{9.17b}$$

where y_n is the target value of sample n, and y_n^{knn} is a vector of targets of the k nearest neighbors of sample n, including a bias term of 1. In addition, w^{knn} is a vector of weights, and N is the number of samples in the training set. The cost function (9.17a) is based on the Mean Absolute Error (MAE) function, which allows point forecasts. Constraint (9.17b) guarantees that all neighbors and the bias term do not contribute negatively to the result of the linear combination.

Optimization problem (9.17) is a proposed extension to the standard KNN algorithm such that quantile forecasts can be obtained by replacing the MAE function with the pinball loss function. Concretely, the optimization problem behind the proposed quantile KNN is structured as follows:

$$J_{\text{knn},q} = \min_{w^{\text{knn}}} \quad \frac{1}{N}\sum_{n=1}^{N} \begin{cases} ((w^{\text{knn}})^{\intercal} y_n^{\text{knn}} - y_n) \cdot q & \text{if } y_n \geqslant (w^{\text{knn}})^{\intercal} y_n^{\text{knn}}, \\ ((w^{\text{knn}})^{\intercal} y_n^{\text{knn}} - y_n) \cdot (1-q) & \text{if } y_n < (w^{\text{knn}})^{\intercal} y_n^{\text{knn}}, \end{cases} \tag{9.18a}$$

$$\text{s.t.} \quad w_i^{\text{knn}} \geqslant 0, \quad \forall i \in \{0, \dots, k\}. \tag{9.18b}$$

9.4.2 *Performance Evaluation*

The performance of both aforementioned quantile forecasting algorithms is evaluated in a real-world LV grid. The analysis also considers multiple levels of PV and EV penetration, which is expected to add even more uncertainty in the future. The concept of short-term probabilistic state forecasting refers to the prediction of the net active and reactive power consumption P_{cons} and Q_{cons}, active and reactive power flow P_{flow} and Q_{flow}, and bus voltage magnitude V with a forecasting horizon of one hour. The set of input features is specific to each quantity. The forecasting accuracy is assessed via the the notions of reliability and sharpness.

9.4.2.1 *Case Study*

The proposed case study is based on the residential LV grid presented in Figure 3.5, which is part of the distribution grid of the City of Basel. As a reminder, only 55% of the end-users were equipped with a smart meter at the time of data preparation. The yearly energy consumption being known for the remaining end-consumers, they are assigned power profiles of smart metered consumers with similar consumption located in other residential areas of Basel. In addition, reactive power pseudo-measurements are created according to the adaptive power factor approach detailed in Section 6.2.3.3. Since reactive power injections are synthetic data, the performance evaluation principally focuses on the other quantities. Together with the voltage measurement at the transformer, an observable grid can be achieved. As detailed in the following, PV production and EV consumption profiles are further added to the base load in order to assess their impact on the grid state uncertainty.

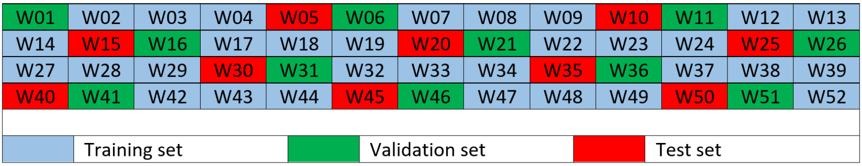
Concretely, a situation where up to 60% of the houses are covered with photovoltaic panels is simulated. For that purpose, the power output profiles of 116 actual PV systems spread in the entire City of Basel are first selected and normalized by their maximal power value. Second, they are suitably allocated to the houses in the case study based on a tool developed by UVEK [385] and Energie Schweiz [386] which assesses the solar potential of any Swiss rooftop. Furthermore, weather data are provided by a meteorological station of MeteoSwiss in the City of Basel [237]. Due to the absence of weather forecasts, this work assumes a perfect weather forecast, which is a strong assumption. In reality, the uncertainty with respect to PV injection is expected to increase because of inevitable weather forecast errors. Accounting for the fact that this work focuses on hour-ahead forecasting, the presence of large weather forecast errors should nevertheless be limited.

The consumption profiles of EV chargers are derived from the open data set of the “My Electric Avenue” project [387]. In this project, the driving and charging patterns of more than 200 Nissan Leaf vehicles have been recorded in the United Kingdom over 18 months. After data cleaning and filtering, 180 charging profiles at 3.7 kW nominal power are extracted, which corresponds to 30% of the households in the considered grid. Since future home chargers are expected to work mainly between 7.4 kW and 11.1 kW [388], the charging power is scaled up while the charging time is accordingly reduced in order to keep the same energy consumption. Electric vehicles associated with a 7.4 kW charger are modeled with the same energy consumption as Nissan Leaf vehicles, whereas EVs with a 11.1 kW charger are assumed to consume similarly as the Audi e-tron or Tesla models, and their energy consumption is multiplied by 1.8 [389]. Finally, charging profiles are allocated to the grid buses according to the algorithm presented in [390] which creates EV clusters. In fact, the algorithm reflects the social effect of increased willingness to purchase an EV when the neighbors also drive an EV.

All aforementioned measurement data are adjusted to a temporal resolution of 15 minutes and limited to a period of one year. In order to represent different DER penetration levels, multiple data sets are created by adding an increasing number of the EV power consumption and PV production profiles to the initial load profiles. This is summarized in Table 9.2 which also indicates the share of houses whose rooftop is equipped with PV panels and the share of households in possession of an EV. In addition, the share of 11.1 kW chargers is increased with the DER penetration to simulate a

Data set	Number of PVs (share of houses)	Number of EVs (share of households)	Share of 11.1 kW chargers
DS_B	0	0	-
DS_1	39 (20%)	60 (10%)	50%
DS_2	77 (40%)	120 (20%)	62.5%
DS_3	116 (60%)	180 (30%)	75%

TABLE 9.2: Overview of the different DER modified data sets used in the case study.



🟩 W01	🟦 W02	🟦 W03	🟦 W04	🟥 W05	🟩 W06	🟦 W07	🟦 W08	🟦 W09	🟥 W10	🟩 W11	🟦 W12	🟦 W13
🟦 W14	🟥 W15	🟩 W16	🟦 W17	🟦 W18	🟦 W19	🟥 W20	🟩 W21	🟦 W22	🟦 W23	🟦 W24	🟥 W25	🟩 W26
🟦 W27	🟦 W28	🟦 W29	🟥 W30	🟩 W31	🟦 W32	🟦 W33	🟦 W34	🟥 W35	🟩 W36	🟦 W37	🟦 W38	🟦 W39
🟥 W40	🟩 W41	🟦 W42	🟦 W43	🟦 W44	🟥 W45	🟩 W46	🟦 W47	🟦 W48	🟦 W49	🟥 W50	🟩 W51	🟦 W52

🟦	Training set	🟩	Validation set	🟥	Test set

FIGURE 9.10: Weeks of the data set split into training set, validation set, and test set.

probable decrease of the price gap between 7.4 kW and 11.1 kW chargers. Subsequently, all bus voltages and power line flows are determined by load flow simulations to complete the system state of the four penetration scenarios. Finally, the four data sets are split into training, validation, and test sets according to Figure 9.10. This partitioning accounts for the seasonal behavior of measurements while still yielding continuous test sets.

9.4.2.2 *Feature Set*

Tables 9.3 and 9.4 summarize the basic sets of input features for bus quantities (i.e., active and reactive net power consumptions, and voltage magnitudes) and line quantities (i.e., active and reactive power flows), respectively. Note that the forecast of a certain quantity also leverages measurements of other quantities to profit from their physical coupling in the grid. For example, the prediction of power line flows benefits from the knowledge of voltages at both ends. All continuous features are scaled by min-max normalization between zero and one, and all categorical variables (i.e., calendar features) are one-hot encoded. One-hot encoding transforms each categorical variable into a vector of binary variables, which ensures that higher values are not considered as

Category	Features	Time delay
Online	V, P_{cons} and Q_{cons}	-
Recent	V, P_{cons} and Q_{cons}	1 hour and 2 hours
Historical	V, P_{cons} and Q_{cons}	1 day and 2 days
Weather	temperature and solar irradiance	-
Calendar	hour, weekday and holiday flag	-

Table 9.3: Basic feature set for the probabilistic prediction of bus quantities.

Category	Features	Time delay
Online	V_1, V_2, P_{flow} and Q_{flow}	-
Recent	V_1, V_2, P_{flow} and Q_{flow}	1 hour and 2 hours
Historical	V_1, V_2, P_{flow} and Q_{flow}	1 day and 2 days
Weather	temperature and solar irradiance	-
Calendar	hour, weekday and holiday flag	-

Table 9.4: Basic feature set for the probabilistic prediction of line quantities.

more important by the forecasting algorithm. Furthermore, Lasso regression and Principal Component Analysis (PCA) have been tested to reduce the feature set dimension. None of the methods appears to be conclusive for the quantile NN in terms of forecast accuracy. However, the Lasso regression enables a reduction of a few percent of the prediction error induced by the quantile KNN regarding the voltage magnitude and active power quantities.

This study investigates the added value of real-time SM data compared to only time-delayed SM data. In the latter case, only day-ahead SM measurements up to midnight are assumed to be accessible. Hence, online and recent features are not available, but the historical feature set is further enhanced by neighboring values around the corresponding values one and two days in the past. This ensures a sufficient number of grid measurements in the feature set. Moreover, the impact of knowing the starting time and duration of EV charging events is also assessed. In practice, this could be inferred from the GPS location and the state of charge of the vehicle[4]. In this case, a binary EV charging feature is added to the buses associated with an EV charger and to all lines connected to those buses.

---

[4] This raises obvious privacy concerns but is out of the scope of this study.

9.4.2.3 *Probabilistic Evaluation Metrics*

The performance of probabilistic forecasting algorithms is evaluated according to their ability to properly estimate the prediction uncertainty, as proposed by the authors in [391]. Concretely, two quantile forecasts build a certain Prediction Interval (PI) associated with a certain confidence level. For example, quantile forecasts with $q = 0.1$ and $q = 0.9$ build a PI with 80% nominal confidence. On this basis, the notions of reliability (REL), Average Coverage Error (ACE), and Average Interval Score (AIS) are defined as follows [391]:

$$\mathrm{REL}(B_Q) = \frac{100\%}{T} \sum_{t=1}^{T} \mathbb{1}_{\hat{y}_{t,50-Q/2} \leqslant y_t \leqslant \hat{y}_{t,50+Q/2}}, \tag{9.19a}$$

$$\mathrm{ACE}(B_Q) = \mathrm{REL}(B_Q) - Q, \tag{9.19b}$$

$$\mathrm{AIS}(B_Q) = \sum_{t=1}^{T} \left(\hat{y}_{t,50+Q/2} - \hat{y}_{t,50-Q/2} \right), \tag{9.19c}$$

where B_Q is the predefined PI (or quantile band), y_t is the target value at time step t, and $\hat{y}_{t,50-Q/2}$ and $\hat{y}_{t,50+Q/2}$ are the quantile forecasts defining the lower and upper bounds of B_Q at time t, respectively. In addition, $\mathbb{1}_z$ is the indicator function under condition z, Q is the nominal confidence of B_Q, and T is the number of time steps in the test set. The reliability represents the percentage of targets that can be effectively captured within the predefined PI. The ACE evaluates whether the reliability is in line with the PI nominal confidence (i.e., Q). The AIS is a measure of the sharpness (or width) of the PI. In this context, a good quantile forecast implies both a good sharpness (i.e., narrow PI) and good reliability (i.e., close to the PI nominal confidence). In other words, the forecasting algorithm aims for a low AIS while keeping the ACE close to zero. If the ACE is positive (or negative), the PI encompasses too many (or too few) target values. Note that the widely used ranked probability score, known as the probabilistic counterpart of the RMSE, is not considered in this study. It cannot correctly capture the poor performance of quantile forecasts when they all overestimate the target value.

9.4.2.4 *Results*

In this section, the outcome of the aforementioned quantile forecasting algorithms is evaluated in different conditions, notably accounting for different levels of EV and PV penetration. Time-series visualization offers a first qualitative insight into the probabilistic forecasting performance. In the second stage, the ACE and AIS allow for quantitative assessment of the reliability and sharpness of the PIs. The evaluation focuses on power flow and bus voltage forecasts which are relevant quantities for DSOs.

Figure 9.11 shows the resulting quantile forecasts of the NN for active power consumption over three days at a specific bus with PV panels and EVs in the DER penetration scenario DS_3. First of all, the large PV injection during the first day and the EV consumption in the first two evenings are noticeable. Most uncertainty appears to come from car charging events such that the additional car charging feature allows for a drastic drop in the forecast uncertainty. The availability of online SM measurements helps to forecast the volatile base household loads and to reduce the uncertainty associated with the PV injection. It also enables the detection of EV charging events with a one-hour time delay if the car charging feature is not provided.

Furthermore, Figure 9.12 illustrates the quantile forecasts based on both algorithms for active power flow on a specific line over the same three days. Whereas the load of EVs, charging at different time periods, is moderate, the simultaneous power injection of multiple PV systems is clearly visible. The NN algorithm can reasonably forecast the power flow and properly detect the periods with higher uncertainty, even without online SM measurements. However, the KNN algorithm produces very narrow prediction intervals which fail to encompass the target values, except during high PV injection time. This phenomenon is observed for each forecast where the difference between the minimum and maximum target values is relatively large. The reason lies in the nature of the KNN algorithm, where the quantile predictions are linear combinations of previous target values. Since the optimal weights defined by Equation (9.18) are applied to all time steps, the prediction intervals get narrower when the target values get closer to zero.

Concerning voltage magnitude forecasts, Figure 9.13 compares the outcome of the NN and KNN algorithms with and without online SM measurements, again in the highest DER penetration scenario over the same three days. All variants are relatively accurate. The shape of the voltage profile is barely impacted by the EV load. In contrast, voltage values largely exceed the overvoltage limit during sunny days, whereas the detection of an overvoltage depends on the considered quantile on cloudy days. In this case, the accuracy of the quantile forecasts is determinant when used by voltage control strategies as presented in Section 9.5. While still being reliable, the KNN algorithm tends to produce narrower prediction intervals than the NN. Reactive power quantities are not explicitly visualized since they do not come from direct measurements. Nevertheless, they appear particularly volatile in this case study, which makes them hardly predictable.

Figures 9.14 and 9.15 evaluate the reliability and sharpness, respectively, of both considered algorithms for line power flows. The outcome focuses on prediction intervals with nominal confidence levels of 50% and 80%. It is presented under the form of box and whisker plots, where each data point represents the metric for a single line or a single bus. The central bar indicates the median value, the small red square is the mean value, the box corresponds to the Interquartile Range (IQR), and the ends of the whiskers define $1.5 \times \text{IQR}$ below and above the lower and upper quartiles, respectively. First of all, substantial differences appear between both prediction algorithms. As noticed in Figure 9.12, the prediction intervals obtained by the KNN algorithm are narrower, which leads to a lower AIS in comparison with the NN algorithm. By definition, the AIS increases with larger PI nominal confidence levels. However, the excellent sharpness of the KNN comes at the cost of poor reliability. In this case, the algorithm creates too narrow prediction intervals that miss the target values much more often than defined by the nominal confidence level, which is not acceptable. In contrast, the NN exhibits excellent reliability and appears well suited for probabilistic power forecasting. Furthermore, the AIS for the NN algorithm increases with rising PV and EV penetration. This indicates that DERs bring more uncertainty to the system. Note also that both metrics are barely impacted by the availability of online SM measurements and by the car feature. Similar outcomes can be observed for the prediction of net active power consumption and are therefore not explicitly analyzed in this section.

Finally, the performance of quantile voltage forecasts is illustrated by Figures 9.16 and 9.17. Both algorithms are characterized by excellent reliability (i.e., ACE close to zero) on average, although the variance among different buses is more significant for the NN algorithm, especially for PIs with 50% nominal confidence. In terms of sharpness, a substantial increase of the AIS is visible from 50% to 80% nominal confidence levels. A somewhat lower score can also be seen for the KNN in comparison with the NN algorithm. Since all voltage values lie in the same range (i.e., around 1 pu), the narrower prediction intervals produced by the KNN algorithm still properly envelop the target values. Based on the linear combination of similar grid states, the KNN algorithm is appropriate for probabilistic voltage forecasting, in contrast to power forecasting. Moreover, online SM measurements allow for slightly narrower prediction intervals.

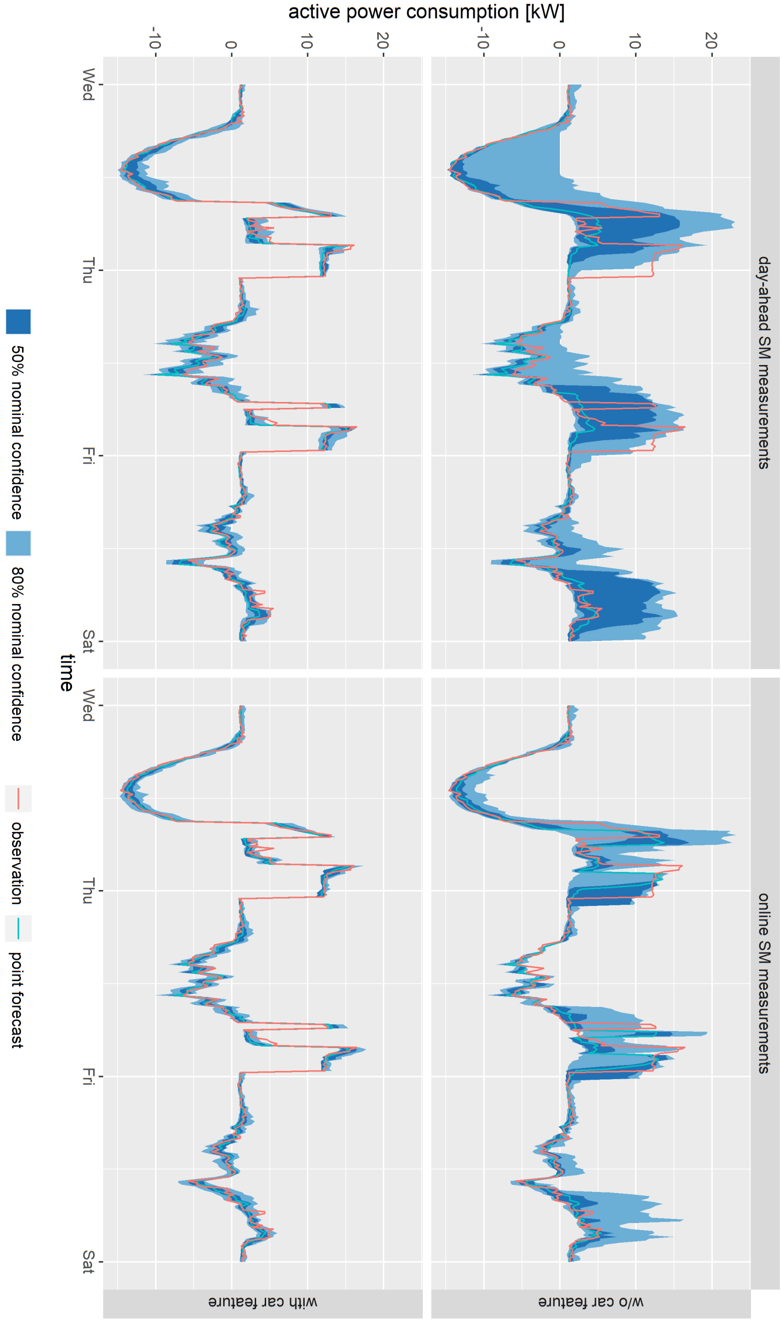


FIGURE 9.11: Quantile forecasts of active power consumption performed by the neural network at a specific bus with PV panels and EVs for scenario DS_3. Prediction intervals with 50% and 80% are visualized.

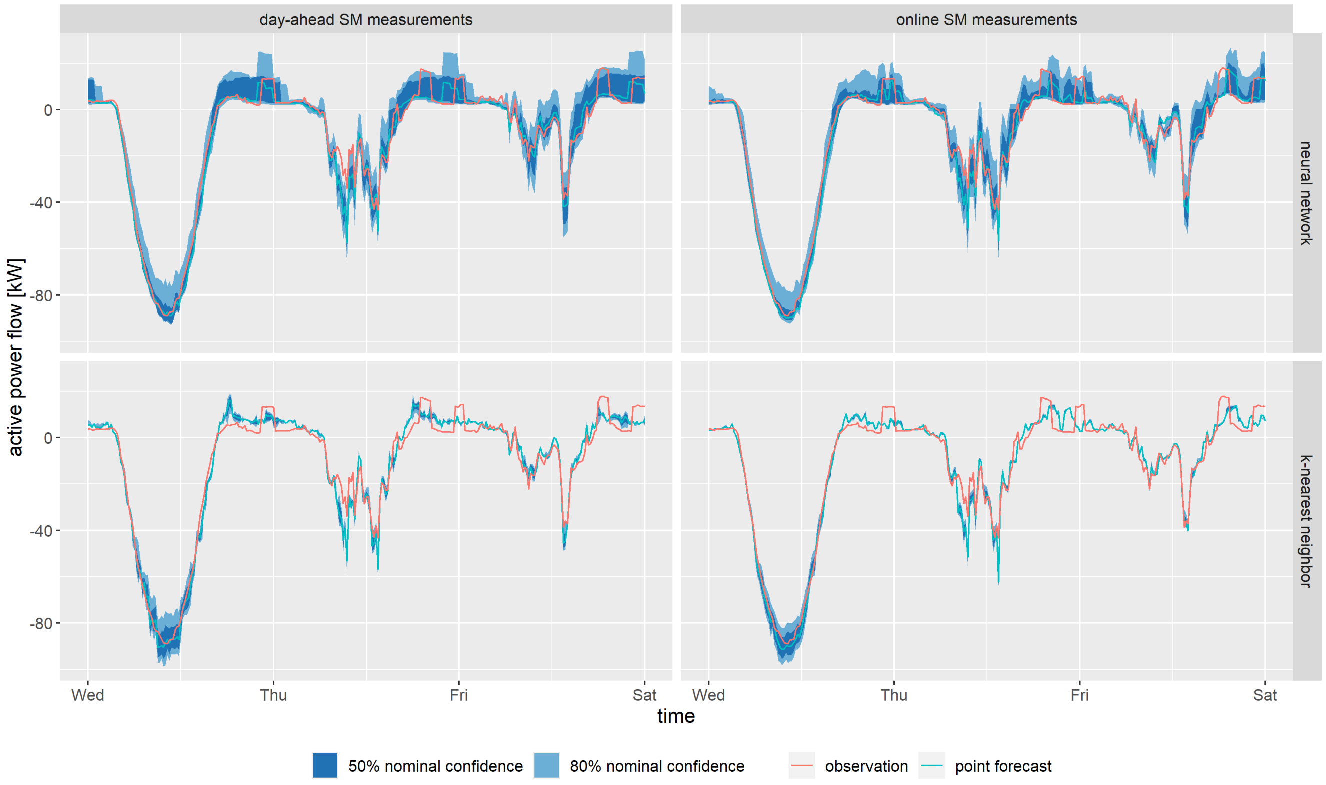


FIGURE 9.12: Quantile forecasts of active power flow at a specific line for scenario DS_3, accounting the car feature. Prediction intervals with 50% and 80% nominal confidence are visualized.

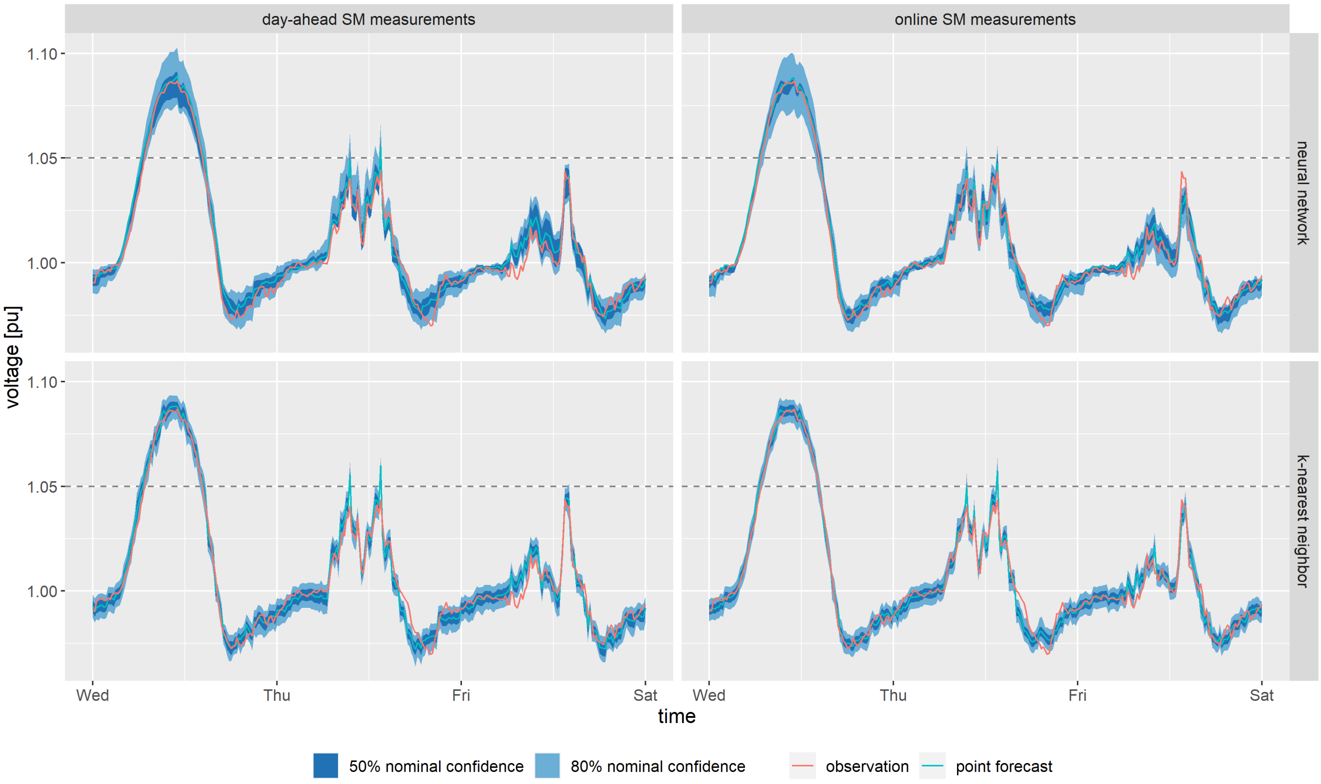


FIGURE 9.13: Quantile forecasts of voltage magnitude at a specific bus with PV panels and EVs for scenario DS_3, accounting for the car feature. Prediction intervals with 50% and 80% nominal confidence are visualized.

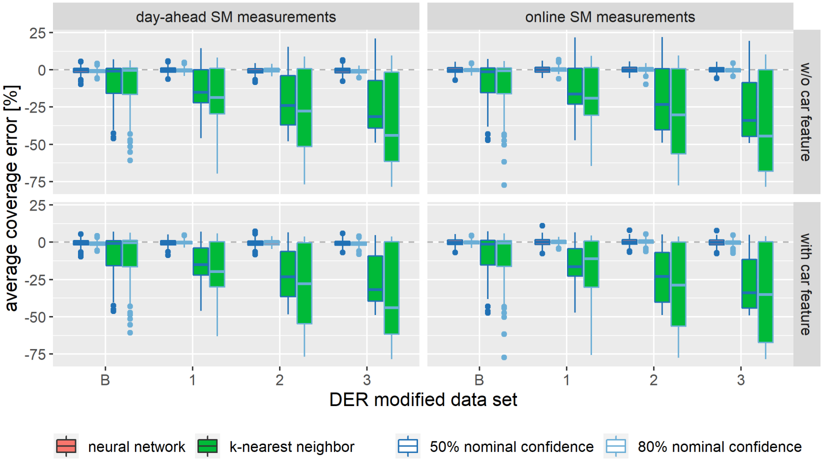


FIGURE 9.14: Average coverage error of the quantile forecasts of active power flow for all lines and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered. [5]

9.5 PREVENTIVE VOLTAGE CONTROL UNDER UNCERTAINTY

As shown by the results in Section 9.4.2.4, overvoltages around noon due to large and simultaneous PV injections are expected to be a major problem in future LV grids. Reactive power control and active power curtailment are the main control measures to mitigate the impact of PV injection on the system [392]. Reactive power control is a cost-effective means to regulate the voltage, but its forecast is associated with considerably large uncertainty. In order to demonstrate the added value of quantile forecasts, this work relies on active power curtailment to keep the voltages below an upper limit. PV active power curtailment simultaneously reduces potential line overloadings.

In that respect, this section presents an approach that aims to optimize the physical DSO position on the intra-day market. Concretely, if an overvoltage can be estimated one hour ahead, an optimization problem computes the

---

[5] Note that the red fill characterizing the NN is barely visible due to the narrow IQRs.

[6] Note that the green fill characterizing the KNN is barely visible due to the narrow IQRs.

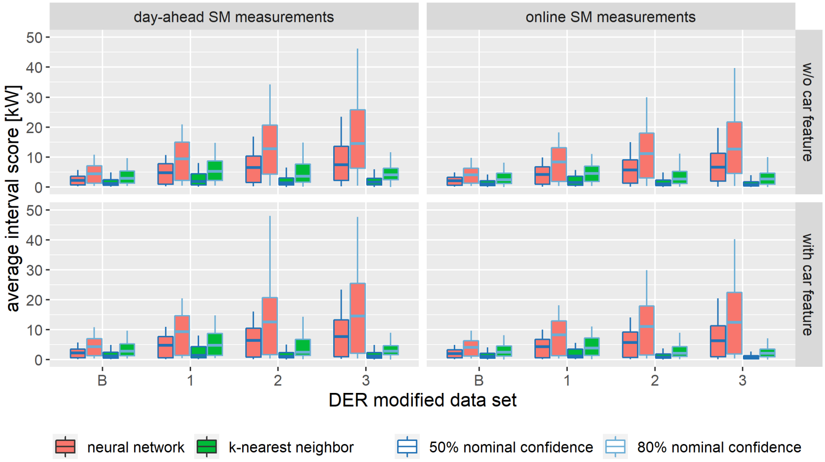


FIGURE 9.15: Average interval score of the quantile forecasts of active power flow for all lines and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered.

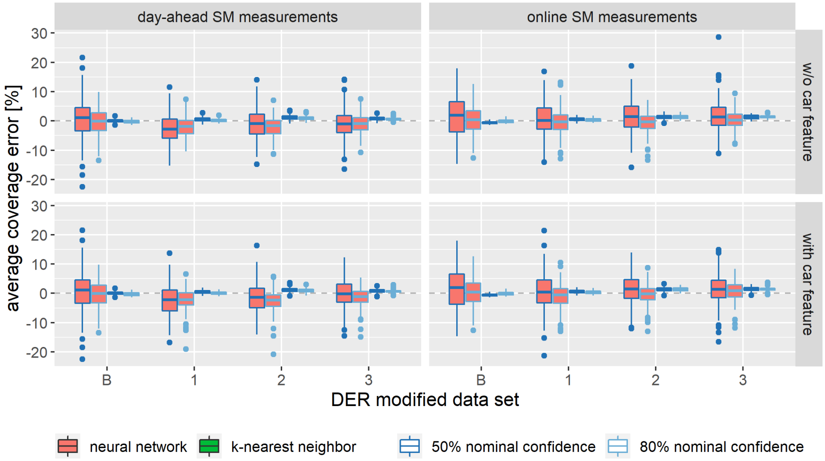


FIGURE 9.16: Average coverage error of the quantile forecasts of voltage magnitude for all buses and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered. [6]

optimal level of PV curtailment with respect to the future estimated state in

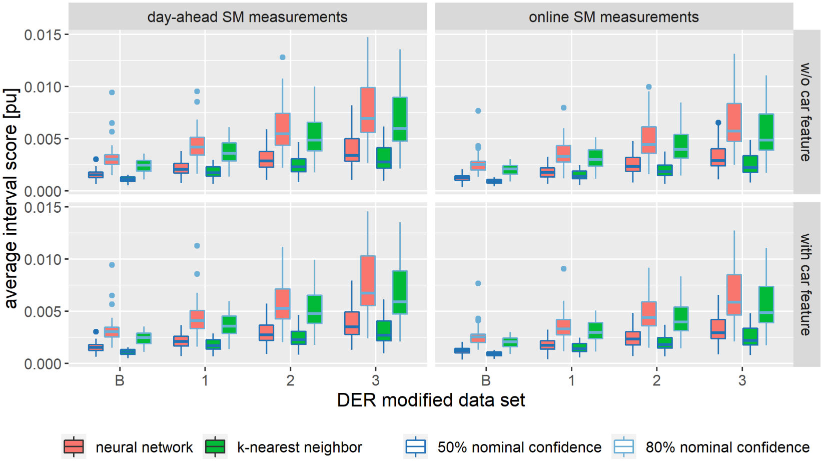


Figure 9.17: Average interval score of the quantile forecasts of voltage magnitude for all buses and time steps in the test set. Prediction intervals with 50% and 80% nominal confidence are considered.

order to comply with the voltage constraints. In this case, it is assumed that the DSO must only compensate the owners of the curtailed PV systems based on the corresponding market price. If an overvoltage is not (fully) eliminated in advance, the DSO has to further curtail PV energy in real-time, which is penalized by a higher imbalance price [393]. This results in a trade-off between the risk of curtailing too much energy in advance and facing a higher price for potential adjustments in real-time. In this section, a pure real-time optimization strategy is compared with control strategies based on point and quantile forecasts. While still widely used in the literature on the control of active distribution grids, perfect forecasts are unrealistic and therefore not considered in this study. Section 9.5.1 presents the different optimization strategies, which are subsequently compared in a real-world case study in Section 9.5.2.

9.5.1 *Methodology*

The different optimization problems for voltage control in an LV grid are designed as AC-OPF problems. They are only solved in case an overvoltage is observed or predicted, where the upper voltage limit is set to 1.05 pu. The optimization variables are the level of PV active power curtailment for each

PV system. In addition, the grid topology is assumed to be perfectly known, the loads and PV systems are assumed to be voltage-independent, and the transformer voltage is assumed to be maintained after the optimization. In the following, a benchmark real-time optimization strategy, an optimization strategy using point forecasts, and an optimization strategy using quantile forecasts are defined. These are referred to as S_B, S_{50}, and S_q, respectively. Each of the latter two optimization strategies consists of two subsequent optimization problems. The first optimization problem is solved one hour before actual PV curtailment, whereas the second optimization problem occurs in real-time, if necessary.

9.5.1.1 *Benchmark Real-Time Optimization Strategy*

In the benchmark strategy S_B, PV curtailment is applied in real-time for each time step subject to an overvoltage, i.e.,

$$\max_{k \in \Psi_\mathrm{B}} V_k > 1.05, \tag{9.20}$$

where V_k is the voltage magnitude at bus k, and Ψ_B is the set of all buses.

In this case, PV curtailment is associated with a high imbalance price and the optimization problem is defined as:

$$\min_{P_{n_\mathrm{PV}}^\mathrm{curt}} \quad \sum_{\Psi_\mathrm{PV}} \frac{1}{4} \cdot C_\mathrm{ib} \cdot P_{n_\mathrm{PV}}^\mathrm{curt}, \tag{9.21a}$$

$$\text{s.t.} \quad 0 \leqslant P_{n_\mathrm{PV}}^\mathrm{curt} \leqslant P_{n_\mathrm{PV}}^\mathrm{prod}, \quad \forall n_\mathrm{PV} \in \Psi_\mathrm{PV}, \tag{9.21b}$$

$$P_k + \sum_{\Psi_{\mathrm{PV},k}} P_{n_\mathrm{PV}}^\mathrm{curt} + \sum_{m \in \Omega_k} P_{km} = 0, \quad \forall k \in \Psi_\mathrm{B}, \tag{9.21c}$$

$$\begin{aligned} P_{km} &= (V_k)^2 \cdot g_{km} \\ &\quad - V_k \cdot V_m \cdot (g_{km} \cdot \cos(\theta_{km}) + b_{km} \cdot \sin(\theta_{km})), \quad \forall k, m \in \Psi_\mathrm{B}, \end{aligned} \tag{9.21d}$$

$$Q_k + \sum_{m \in \Omega_k} Q_{km} = 0, \quad \forall k \in \Psi_\mathrm{B}, \tag{9.21e}$$

$$\begin{aligned} Q_{km} &= -(V_k)^2 \cdot b_{km} \\ &\quad + V_k \cdot V_m \cdot (b_{km} \cdot \cos(\theta_{km}) - g_{km} \cdot \sin(\theta_{km})), \quad \forall k, m \in \Psi_\mathrm{B}, \end{aligned} \tag{9.21f}$$

$$0.95 \leqslant V_k \leqslant 1.05, \quad \forall k \in \Psi_\mathrm{B}, \tag{9.21g}$$

$$V_1 = V_1^\mathrm{meas}, \tag{9.21h}$$

$$\theta_1 = 0, \tag{9.21i}$$

where $P_{n_\mathrm{PV}}^\mathrm{curt}$ and $P_{n_\mathrm{PV}}^\mathrm{prod}$ are the curtailed power and the production potential of PV system n_PV, respectively. P_k and Q_k are the net active and reactive power consumptions before curtailment at bus k, respectively. P_{km} and Q_{km} are the active and reactive power flows from bus k to bus m, respectively, and V_1^{meas} is the voltage magnitude measured at the slack bus. Parameters g_{km} and b_{km} are the line conductance and susceptance from bus k to bus m, θ_{km} is the voltage angle difference between bus k and bus m, and θ_k is the voltage angle at bus k. Ψ_{PV} is the set of all PV systems in the network, $\Psi_{\text{PV},k}$ is the set of all PV systems connected to bus k, Ω_k is the set of all power lines connected to bus k, and C_{ib} is the imbalance price of curtailed PV energy. The cost function (9.21a) defines the total curtailment cost at the time step under consideration[7]. Constraint (9.21b) limits the PV power curtailment between 0 and the total potential PV production. The active and reactive node balances as well as the AC power flow equations are defined in (9.21c)–(9.21f). Constraint (9.21g) ensures that the voltage stays within the predefined limits of $1 \pm 5\%$ pu, and constraints (9.21h) and (9.21i) set the voltage magnitude and angle reference at the slack bus, respectively.

9.5.1.2 *Optimization Strategy using Point Forecasts*

Assuming that point forecasts (i.e., 50%-quantile forecasts) are available, strategy S_{50} is used for each time step where an overvoltage is forecasted one hour ahead, i.e.,

$$\max_{k \in \Psi_{\text{B}}} \hat{V}_{50,k} > 1.05, \tag{9.22}$$

where $\hat{V}_{50,k}$ is the point forecast of the voltage magnitude at bus k.

In a first stage, point forecasts of the net power consumptions are integrated into the following optimization problem:

$$\min_{P_{n_{\text{PV}}}^{\text{curtHA}}} \quad \sum_{\Psi_{\text{PV}}} \frac{1}{4} \cdot C_{\text{m}} \cdot P_{n_{\text{PV}}}^{\text{curtHA}}, \tag{9.23a}$$

$$\text{s.t.} \quad 0 \leqslant P_{n_{\text{PV}}}^{\text{curtHA}} \leqslant \hat{P}_{n_{\text{PV}}}^{\text{prod}}, \quad \forall n_{\text{PV}} \in \Psi_{\text{PV}}, \tag{9.23b}$$

$$\hat{P}_{50,k} + \sum_{\Psi_{\text{PV},k}} P_{n_{\text{PV}}}^{\text{curtHA}} + \sum_{m \in \Omega_k} P_{km} = 0, \quad \forall k \in \Psi_{\text{B}}, \tag{9.23c}$$

$$\begin{aligned} P_{km} &= (V_k)^2 \cdot g_{km} \\ &\quad - V_k \cdot V_m \cdot (g_{km} \cdot \cos(\theta_{km}) + b_{km} \cdot \sin(\theta_{km})), \quad \forall k, m \in \Psi_{\text{B}}, \end{aligned} \tag{9.23d}$$

$$\hat{Q}_{50,k} + \sum_{m \in \Omega_k} Q_{km} = 0, \quad \forall k \in \Psi_{\text{B}}, \tag{9.23e}$$

$$Q_{km} = -(V_k)^2 \cdot b_{km} \tag{9.23f}$$

---

[7] Factor $\frac{1}{4}$ accounts for a temporal resolution of 15 minutes.

$$+ V_k \cdot V_m \cdot (b_{km} \cdot \cos(\theta_{km}) - g_{km} \cdot \sin(\theta_{km})), \quad \forall k, m \in \Psi_{\mathrm{B}},$$

$$0.95 \leqslant V_k \leqslant 1.05, \quad \forall k \in \Psi_{\mathrm{B}}, \tag{9.23g}$$

$$V_1 = \hat{V}_{50,1}, \tag{9.23h}$$

$$\theta_1 = 0, \tag{9.23i}$$

where $\hat{P}^{\mathrm{prod}}_{n_{\mathrm{PV}}}$ is the point forecast of the potential power production of PV system n_{PV}, and $P^{\mathrm{curtHA}}_{n_{\mathrm{PV}}}$ is the hour-ahead curtailed power of PV system n_{PV}. $\hat{P}_{50,k}$ and $\hat{Q}_{50,k}$ are point forecasts of the active and reactive power consumption at bus k, respectively, and C_{m} is the market price of curtailed PV energy[8]. Note that separate point forecasts are performed for the potential power production of each PV system. This optimization problem is similar to Problem (9.21) of the benchmark strategy S_B at the exception that hour-ahead point forecasts replace the realizations and that the marginal cost associated with PV curtailment is the market price which is lower than the imbalance price. At this stage, the DSO enforces PV curtailment for the time step under consideration as defined by the optimization problem.

In a second stage, the PV power injection is further adjusted in real-time whenever remaining overvoltages are observed. Concretely, a similar optimization problem as Problem (9.21) of the benchmark strategy S_B is solved whenever Condition (9.20) is satisfied. Nevertheless, the decisions made one hour before must be included by replacing Equations (9.21b) and (9.21c) with the following constraints:

$$0 \leqslant P^{\mathrm{curt}}_{n_{\mathrm{PV}}} \leqslant P^{\mathrm{prod}}_{n_{\mathrm{PV}}} - P^{\mathrm{curtHA}}_{n_{\mathrm{PV}}}, \quad \forall n_{\mathrm{PV}} \in \Psi_{\mathrm{PV}}, \tag{9.24a}$$

$$P_k + \sum_{\Psi_{\mathrm{PV},k}} P^{\mathrm{curtHA}}_{n_{\mathrm{PV}}} + \sum_{\Psi_{\mathrm{PV},k}} P^{\mathrm{curt}}_{n_{\mathrm{PV}}} + \sum_{m \in \Omega_k} P_{km} = 0, \quad \forall k \in \Psi_{\mathrm{B}}. \tag{9.24b}$$

9.5.1.3 *Optimization Strategy using Quantile Forecasts*

In a first stage, strategy S_q leverages quantile voltage forecasts to minimize the cost of overvoltages. For this purpose, the voltage uncertainty is defined as the difference between the quantile forecast and the point forecast:

$$d\hat{V}_{q,k} = \hat{V}_{q,k} - \hat{V}_{50,k}, \quad \forall q \in (0, 100), \ \forall k \in \Psi_{\mathrm{B}}, \tag{9.25}$$

where $d\hat{V}_{q,k}$ and $\hat{V}_{q,k}$ are the voltage uncertainty and the point forecast of the voltage for quantile q at bus k, respectively. The curtailment strategy is applied at each time step for which the following condition holds:

$$\max_{k \in \Psi_{\mathrm{B}}} \hat{V}_{q,k} > 1.05. \tag{9.26}$$

---

[8] Remaining variables and parameters are defined in Problem (9.21)

The first-stage optimization problem of strategy S_q only differs from Problem (9.23) of strategy S_{50} in the formulation of the voltage limit defined by Equation (9.23g) and the slack bus voltage defined by Equation (9.23h), which must be replaced with the following constraints:

$$0.95 \leqslant V_k + d\hat{V}_{q,k} \leqslant 1.05, \quad \forall k \in \Psi_{\mathrm{B}}, \tag{9.27a}$$

$$V_1 = \hat{V}_{50,1} - d\hat{V}_{q,1}. \tag{9.27b}$$

Note that the voltage forecasts are only included in the form of the uncertainty of the quantile, whereas the resulting voltages are determined by optimization. In addition, only point forecasts of the power consumption are used in order not to mix the uncertainties from different forecasting sources.

In a second stage, potentially remaining overvoltages detected in real-time are handled by the same second-stage optimization as for strategy S_{50}.

9.5.2 *Results and Discussion*

The different voltage control strategies are applied to the case study defined in Section 9.4.2.1 for scenario DS_3 which simulates a PV and EV penetration of 60% and 30%, respectively. The quantile forecasting algorithms showing the best performance in Section 9.4.2.4 are used for the different forecasts, which corresponds to the NN and the proposed KNN algorithm for power and voltage quantities, respectively. In any case, online SM measurements and the car charging feature are used. Note that power line flows are not directly given since they are implicitly calculated by the OPF problems.

In addition, an average ACE $\approx -6\%$ is observed for the 50%-quantile voltage forecasts at the time steps where an overvoltage is predicted, which indicates that the point forecasts tend to overestimate more than underestimate the voltage at these time steps. This gives the incentive to relax the voltage limit to the 44%-quantile in strategy S_q. Alternatively, due to a lower market price than imbalance price, the DSO might want to remove as much overvoltage as possible at the first stage and even accept superfluous PV curtailment. This would justify the use of a higher quantile in strategy S_q, which is set to 62.5% in this case study.

Moreover, since the results are evaluated on a Swiss grid, the market price C_m for electrical energy is set to 40 €/MWh which roughly corresponds to the average Swiss spot market price [394]. Subsequently, the Swiss imbalance price is defined as [393]:

$$C_{\mathrm{ib}} = 1.1 \cdot (1.2 \cdot C_{\mathrm{m}} + p_{\mathrm{ib}}), \tag{9.28}$$

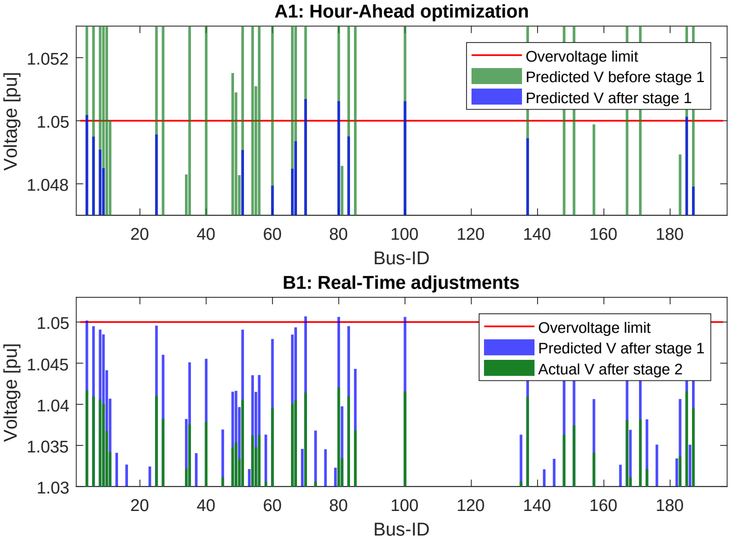


Figure 9.18: Intermediate voltage point forecasts and final voltage realizations resulting from strategy S_{44} at a selected time step.

where C_{m} and C_{ib} are the market and imbalance prices for the curtailed PV energy in €/MWh, respectively, and p_{ib} is an imbalance penalty equal to 10 €/MWh. Considering the increasing share of volatile DERs in future distribution grids, imbalances are expected to increase, which could lead to an increase of the imbalance price with respect to the market price. Hence, a future situation where the imbalance price would be doubled is also considered.

First of all, Figure 9.18 gives insight into the functioning of S_{44} for a given time instance. More precisely, Subplot A1 illustrates the initial voltage forecast and the predicted voltage result applying the first-stage optimization for the 50%-quantile and for all buses in the network. Since this strategy reduces the overvoltages only to the 44%-quantile, there are still remaining overvoltages predicted for the 50%-quantile at a few buses. Subsequently, Subplot B1 shows the predicted voltages after the first-stage optimization and the actual voltage realizations after applying the real-time optimization. The real-time optimization is designed to adjust the curtailment strategy in order to guarantee that all voltages are below the overvoltage limit. At this specific time instance, it nevertheless appears that all the voltages are already way below the limit after applying the first-stage strategy, even

Curtailed energy and cost	S_B	S_{50}	S_{44}	$S_{62.5}$
HA energy curtailment [MWh]	0.0	22.3	21.6	23.7
RT energy curtailment [MWh]	22.3	2.1	2.4	1.5
Total energy curtailment [MWh]	22.3	24.4	24	25.2
Total cost in current situation [€]	1422	1022	1017	1044
Total cost in future situation [€]	2843	1154	1172	1142

TABLE 9.5: Optimization results for different voltage control strategies in current and potentially future price situations, considering the curtailed energy and the associated cost.

though there were reduced only to the 44%-quantile[9]. Consequently, no real-time adjustment is required. Due to the tendency of the point forecast to overestimate overvoltages, strategy S_{44} exactly profits of such scenarios where the overvoltage can be completely eliminated at the hour-ahead stage by curtailing less energy than the point forecast would have predicted.

Finally, Table 9.5 compares the different strategies for both price situations in terms of curtailed energy and cost. The total cost consists of the cost of the Hour-Ahead (HA) and of the Real-Time (RT) curtailments for a period of 10 weeks evenly distributed over a year, as shown in Figure 9.10. Although the forecast-based strategies curtail in total more PV energy than the pure real-time optimization because of the prediction errors, they allow for a clear reduction of the total cost. For example, strategy S_{50} increases by 4.9% the amount of curtailed PV energy with respect to the benchmark strategy but reduces by 28.1% and even 59.4% the final cost based on the current and future price situations, respectively. In the current price situation, S_{44} is the most cost-efficient strategy, where the total cost further drops by 0.5% with respect to S_{50}. Since the imbalance price is only about 50% higher than the market price and the point forecast tends to overestimate the voltage in overvoltage situations, it is preferable to enable some more remaining overvoltages that are handled in real-time. Conversely, if the imbalance price increases, it gets profitable to accept more and even too much curtailment in the first stage. This is shown by the lowest total cost for strategy $S_{62.5}$, i.e., -1% with respect to S_{50}. Note that the actual value of the market and imbalance prices influences only the total cost, not the optimal curtailment.

---

[9] If there were any remaining overvoltages after the first-stage optimization, at least one voltage would be right at the voltage limit after the real-time optimization.

Based on Table 9.5, saving potentials might seem relatively low. Nevertheless, this is only the outcome for a small residential neighborhood with 583 consumers over ten weeks. Extrapolated to the whole city of Basel with about 100'000 households [395] for a time period of one year, IWB could potentially save about 360'000 € when applying preventive instead of corrective voltage control[10]. This yearly saving potential even amounts to approximately 1'500'000 € in case of a doubling of the imbalance price.

9.6 CONCLUSION

Different aspects related to short-term forecasting in LV grids are covered in this chapter. First, the inadequacy of standard deterministic algorithms and evaluation metrics is demonstrated for 24-hour-ahead load forecasting on the basis of 1000 load profiles. At this level, the load is known to be particularly difficult to predict. In terms of evaluation, the standard point-wise metrics and the ramp score give a biased image of the forecasting accuracy for volatile loads due to the double penalty effect, which favors smooth predictions. In contrast, the adjusted error metric rewards the prediction of consumption spikes even with slight displacements in time and mitigates the double penalty effect. Nevertheless, all considered algorithms fail to outperform the simple persistence model according to this adjusted error metric. Besides, traditional deterministic algorithms do not reflect the statistical properties of original load profiles, which has been studied for the ARMAX and SVR models. To different extents, the observations can be generalized to all algorithms focusing on the point-wise accuracy. Nonetheless, the performance evaluation of a forecasting algorithm basically depends on the final application. For example, in the context of voltage control, the point-wise accuracy and the statistical properties of load forecasts are not of high importance as long as the voltage remains within acceptable limits. In this case, the performance evaluation of a load forecast should consider the number of resulting voltage violations. Moreover, although rarely considered in the forecasting literature, the voltage is associated with a relatively lower uncertainty such that its deterministic prediction still provides valuable information. This is confirmed in the proposed preventive voltage control scheme, where the strategy based on point forecasts achieves substantial cost savings in comparison with a purely corrective approach.

---

[10] Note that the purpose of this extrapolation is only to obtain a rough order of magnitude for saving potentials at a city level. It is nevertheless clear that a specific neighborhood cannot representatively reflect the situation at a city level.

Although deterministic forecasting appears convenient in specific cases, LV grids are still subject to large uncertainty that cannot be comprehensively predicted by a single estimate. This uncertainty comes from the traditional load, but also increasingly from PV systems and EV chargers which additionally put the grid infrastructure and its safe operation under pressure. In that respect, probabilistic forecasting approaches seem more appropriate than deterministic algorithms. In this chapter, quantile forecasting is leveraged for predicting the hour-ahead grid state, including net power consumptions, power flows, and voltages. Concretely, a quantile neural network is compared with a proposed quantile version of the KNN algorithm. Characterized by relatively high uncertainty, the former is more effective for power quantities, while the latter better estimates voltages. The evaluation is based on the reliability and the sharpness of prediction intervals. The case study also considers various levels of PV and EV penetration, which noticeably impacts the level of uncertainty. The presence of online SM measurements and the knowledge about the EV charging behavior (i.e., starting time and duration) slightly improve the prediction performance.

Finally, point and quantile forecasts are integrated into preventive voltage control schemes via PV power curtailment. The proposed AC-OPF approaches assume that the level of active power curtailment can be decided in advance at a lower cost, however associated with a certain forecast error. While the preventive control schemes are definitely more cost-efficient than their corrective counterpart, the case study shows that quantile forecasts can reduce the costs even further compared to point forecasts. The exact quantiles to consider and the corresponding cost reduction depend on the price difference between hour-ahead and real-time curtailment. Future work should investigate the optimal quantiles that maximize the cost reduction depending on the price situation. Different quantiles could also be integrated into the same optimization problem and even directly contribute to the cost function instead of acting on the voltage limit constraint. The performance of the suggested control scheme should also be compared with other probabilistic approaches such as stochastic or chance-constrained optimization, where the grid quantities are seen as random variables with a certain probability distribution. In addition, alternative means to control voltages such as reactive power control, online tap changing of transformers, the use of the EV flexibility, and demand response should be considered. The presented OPF approach could also be adapted to congestion management, demand-side management, or any application subject to uncertainty and that can benefit from short-term forecasts.

To summarize, this chapter basically encourages the use of actual short-term forecasts for the operation of distribution grids, notably via preventive control. On the one hand, perfect forecasts are absolutely not realistic, accounting for the large uncertainty inherent to distribution grids. On the other hand, the assumption of a predefined probability distribution of the quantities of interest, which would not depend on temporal and spatial dimensions, is reductive. While point forecasts can be valuable in certain cases, the use of probabilistic forecasts should be preferred. Nevertheless, the principal challenge remains in their effective integration into operational schemes, which definitely merits further consideration.

10

CONCLUSIONS AND OUTLOOK

This chapter summarizes the content of this thesis in Section 10.1, draws the main conclusions in Section 10.2, and suggests various avenues for future work in Section 10.3.

10.1 SUMMARY

Although the Advanced Metering Infrastructure (AMI) is still under development in most distribution grids, the generated data already offers excellent opportunities for power utilities and their customers. Beyond the undeniable practical aspect of automated meter reading, smart meters, but also a wide variety of advanced sensors (e.g., installed at cable distribution cabinets, MV/LV transformers, local substations) enhance the visibility and controllability of the system down to the low-voltage grid. All these sensors are supported by appropriate communication networks and data storage systems. The detection of non-technical losses, better monitoring and situational awareness, energy forecasting, demand-side management, or the creation of transactive energy systems are among the most recurrent applications suggested in the literature. There is nevertheless a large gap between the possibilities offered by the AMI, even when complying with the development objectives, and the general assumptions taken in the literature. Notably, the roll-out targets do not necessarily aim for a full smart meter coverage, the availability of digital grid models is not self-evident, especially at the low-voltage level, and the diverse sets of data and metadata are always prone to errors, inaccuracies, and missing values. Moreover, the output granularity and the recorded quantities of advanced meters are limited to keep the amount of generated data under control. Sub-metering is also rarely realized on a large scale due to its high cost. Besides, end-consumers cannot be considered as rational entities when they are given the possibility to be active participants in the system. More importantly, data protection and privacy concerns prevent the implementation of intrusive applications. This crucial aspect also highly limits data and information sharing between power utilities and the research community, which obligates the latter to use synthetic data and rely on case studies which are sometimes far from reality. Having the chance to rely solely on real-world data, the work proposed in this thesis intends to highlight which are the assumptions and simplifications that can realistically be taken in the development and validation of data-based approaches. It also suggests various processes and methods to effectively leverage the realistic possibilities offered by AMIs and address some of the current challenges in grid operation and planning.

First of all, appropriate preparation of data is a primordial step before their use in future applications. The proposed data preparation pipeline brings raw data into a formatted and standardized form in the first stage and detects and fixes potential signs of bad data quality in the second stage. The process can be standardized to a certain extent, but the exact methodologies used for the different cleaning steps largely depend on the type of data and on its future application. It must also be kept in mind that 80% of the time and resources required for a certain data-based process are used by data preparation, whereas the actual analysis only accounts for the remaining 20%.

Furthermore, the high spatial and temporal resolutions provided by smart meters enable a previously unattainable degree of detail in distribution grids but also lead to the generation of so-called big data that cannot be directly integrated into decision-making processes. A good overview of measurement data at the end-user level necessarily entails a certain complexity reduction and proper visualization at a large-scale level. Unsupervised learning techniques, and especially clustering, add value to smart meter data by grouping and putting into perspective the multiple pieces of information. The extraction of features defines the focus of the analysis, which can go much beyond the sole clustering of end-consumers based on the shape of their load profile. The spatial dimension of smart meters can also be leveraged to visualize the clustering outcomes.

Besides, the temporal averaging effect of low temporal resolutions and spatial aggregation substantially impact the properties of load profiles in distribution grids, which is not the case in transmission networks. Among others, the volatility and peak values of consumption data are considerably reduced, even at the level of distribution transformers. Such alteration of the original load properties must be considered in data-based models.

In addition, the design of control strategies or the study of future scenarios (e.g., increased penetration of DERs) in a specific grid inevitably relies on time series simulations in a certain model of that grid. Their conclusions are influenced by the quality and representativeness of this model. Direct measurements provide some information on the characteristics of the grid but are insufficient to obtain a complete and observable system, especially at the LV level. This requires pseudo-measurements which could consist of synthetic load profiles. An adaptation of traditional Markov chain models is proposed for the synthesis of active power profiles based on existing smart meter data. A binary optimization and a bin packing problem are further developed to select the most suitable load profiles for a given system and allocate them to non-metered end-users, respectively. Both approaches especially take care of not altering the properties of the original load profiles. The adaptive Markov chain model clearly outperforms the standard load allocation to reflect the properties of individual consumers, notably in terms of peak values. By integrating the notion of seasonality and periodicity at different time scales in its design, the adaptive model also outperforms its traditional counterpart at an aggregate level. Reactive power pseudo-measurements are generally neglected in the literature. A synthesis approach is suggested in this work on the basis of actual smart meter data. Nevertheless, the synthesis approach at the individual level does not seem to be important as long as the individual reactive power profiles are scaled to match with aggregate measurements.

Moreover, this work includes a comprehensive sensitivity analysis of the main dimensions influencing the state estimation of distribution grids, including the LV level. The analysis accounts for the synthesis of pseudo-measurements, but also the penetration and type of direct measurements, and the placement of smart meters. It appears that a smart meter penetration of 75% already leads to a satisfying modeling accuracy. At lower smart meter coverage, the type of active power pseudo-measurements is relevant, especially for the estimation of peak values which are underestimated by standard load profiles. The synthesis approach for reactive power does not seem to influence the state estimation outcome, although this merits closer scrutiny. Besides, the strategic placement of smart meters to the largest consumers approximately leads to the same accuracy as randomly placed meters with a 25% higher penetration. Furthermore, the estimation of power flows and voltages clearly benefits from the installation of advanced metering devices at distribution cabinets and transformers.

Eventually, the thesis discusses different data-based applications in low-voltage grids and proposes novel methods based on typically available measurement data to address some operational challenges. First, various approaches to detect and disaggregate the consumption profile of cold appliances and storage water heaters are developed. These domestic devices are characterized by a non-negligible flexibility potential, and their operation is not controlled by end-users, which makes them perfect candidates for demand response purposes via direct load control. In this context, the estimation of their power with a relatively high spatial (i.e., at the device level) and temporal (i.e., sampling period of 1 to 30 minutes) resolutions potentially allows power utilities to design more effective demand response schemes. The general availability of sub-metering data is nevertheless not realistic due to cost and privacy reasons. Hence, multiple unsupervised approaches based on standard smart meter data are successfully developed. Such approaches leveraging data with a sampling period between 1 and 30 minutes are very rarely proposed in the literature and are totally novel for cold appliances and water heaters, to the best of the authors' knowledge. It must be noted that a decrease in the temporal resolution strongly affects the disaggregation performance. In the presented case study with a set of 70 households, the disaggregated cold appliances and water heaters account for more than 40% of the peak load, which prefigures the potential for demand response.

Furthermore, the application of short-term forecasting is investigated in low-voltage grids. First, standard deterministic algorithms and evaluation metrics for 24-hour-ahead load forecasting appear inadequate due to the high volatility and low predictability. The double penalty effect is also pointed out for point-wise metrics. According to an adjusted error metric, none of the considered algorithms is able to clearly outperform the simple persistence model. In addition, the statistical properties of the outcome of traditional algorithms are far from reality. Probabilistic forecasting approaches are generally better suited by additionally estimating the uncertainty associated with the point forecast. Quantile forecasting algorithms are therefore applied to the hour-ahead prediction of the LV grid's state (i.e., net power consumption, power flow, voltage). A quantile version of the KNN algorithm is proposed, which outperforms the competing NN for voltage forecasting. Generally, the uncertainty increases with rising EV and PV penetration. Finally, the point and quantile forecasts are integrated into preventive voltage control schemes based on PV power curtailment with promising results.

10.2 CONCLUSIONS

The main conclusions and findings of this thesis are summarized as follows:

- The use of real-world data for distribution grid applications definitely allows for the design of more realistic approaches and case studies and partially bridges the gap with the current scientific literature. For that purpose, closer collaboration between power companies and the scientific community is an absolute prerequisite, without however neglecting data privacy concerns.

- Data preparation is a crucial and non-negligible preliminary step before performing any analysis based on real-world measurements. In fact, there is no one-size-fits-all solution due to the large variety of data sources, and some experience and domain knowledge is inevitably required. The exact formatting and cleaning methodologies primarily depend on the type of data, the application, and the computational resources. In any case, the preparation process influences the subsequent data-based analysis and should be specified together with the analysis setup.
- Accounting for the variety of end-users in distribution grids, simple unsupervised methods (e.g., clustering) and appropriate visualization at the overall system level already offer a good overview of the smart meter data potential for a wide range of applications. Clustering analysis is of interest to identify the main types of consumers but also detect specific groups of loads whose behavior might be of particular interest and merits closer scrutiny. For example, focusing on the correlation of the load with temperature, combined with the spatial dimension of smart meters, can indicate sections of the grid that are potentially sensitive to demand response.
- The temporal averaging effect and spatial aggregation substantially alter the inherent properties of load profiles in distribution grids, which might lead to biased conclusions if there are not explicitly considered in the data-based processes. This concerns models which are impacted by the data volatility (e.g., load disaggregation, forecasting) and approaches that rely on aggregate data for the modeling of data at lower aggregation levels (e.g., pseudo-measurement synthesis).
- The creation of synthetic data is often inspired by proven approaches at the transmission level, but their application to distribution grids is unrealistic, notably due to the highly volatile nature of the load. Hence, an adaptive Markov chain model is proposed for the generation of realistic active power profiles based on smart meter data. It is designed to integrate any features of interest (e.g., statistical properties, temporal dimensions, correlation with exogenous variables) without

an extensive amount of training data and computational resources. It can be leveraged for pseudo-measurement synthesis, missing values imputation, and load forecasting.

- The sole use of point-wise evaluation metrics gives a biased image of the performance of most approaches developed for distribution grids applications (e.g., pseudo-measurement synthesis, state estimation, short-term forecasting). In fact, they unfairly favor smooth outcomes, which is not representative of the volatile nature of the load, especially at the low-voltage level. The performance evaluation should integrate alternative metrics such as the presented adjusted error, consider multiple other aspects such as the representation of peak values and statistical properties, and above all adapt to the application.

- An overwhelming majority of non-intrusive load monitoring techniques either rely on sub-metering data or require a sampling frequency of at least 1 Hz. These assumptions generally do not hold true in current low-voltage grids. However, the disaggregation of cold appliance and water heater loads with decent accuracy on the sole basis of standard smart meter data is still possible.

- The application of deterministic load forecasting algorithms in low-voltage systems is questionable, especially with a horizon further than the following time step (e.g., day-ahead or 24-hour-ahead predictions). In fact, the deterministic outcome generally appears much smoother than the measured load profile. In contrast, quantile forecasting algorithms can properly account for the uncertainty associated with the state of LV grids.

- The use of point and quantile forecasts in preventive voltage control schemes allows for considerable cost savings in comparison with purely corrective measures, even accounting for forecast errors.

10.3 OUTLOOK

There are various avenues for future work and possibilities to extend the topics addressed in this thesis. In the following, the most important questions that remain unanswered or require further investigation are pointed out:

- The temporal averaging effect studied in Chapter 5 is expected to affect the dimensioning of grid components and the detection of voltage band violations and overloadings. A deeper analysis of its impact could be performed, e.g., in the framework of the sensitivity analysis of Chapter 7 and for short-term forecasting applications as presented in Chapter 9.

- The synthesis approach proposed in Chapter 6 for reactive power shows a somewhat mixed performance. The use of ML-based algorithms and disaggregation techniques, as well as additional features, should lead to a more representative outcome.
- The added value of more advanced metering devices such as micro-PMUs at the LV level, but also additional measured quantities such as the current could be evaluated in the sensitivity analysis of Chapter 7.
- In Chapter 7, the WLS algorithm assumes a Gaussian error distribution, which is generally not valid in LV grids. Alternative algorithms should be tested. Accounting for the large uncertainty induced by the lack of direct measurements, the concept of probabilistic state estimation merits closer scrutiny.
- Since distribution grids are typically unbalanced, the case studies should consider three-phase measurements and three-phase digital grid models, especially in the state estimation of Chapter 7, and the state forecasting and voltage control schemes of Chapter 9.
- This is a strong assumption to consider the grid model as perfectly known, notably in Chapters 7 and 9. Similarly to the measurement data, the inevitable inaccuracies and errors in the grid model must be properly handled.
- Chapter 9 demonstrates that a rising penetration of PVs and EVs in a low-voltage system increases the uncertainty and negatively impacts its state forecasting accuracy. This influence should also be investigated in other domains, such as the sensitivity analysis of Chapter 7.
- The detection and disaggregation approaches presented in Chapter 8 could be adapted to additional flexible loads (e.g., heat pumps) and to online versions. Moreover, beyond the sole estimation of the instantaneous power of these devices, the data-based estimation of their flexibility potential at the device level could greatly benefit demand response schemes.
- The AC-OPF scheme proposed in Chapter 9 for preventive voltage control should be further developed and compared against state-of-the-art stochastic or chance-constraint optimization. More generally, deeper

investigation is required for the effective integration of probabilistic forecasts in preventive control schemes for grid operation purposes.

BIBLIOGRAPHY

[1] *Energy Strategy 2050*. URL: `https://www.bfe.admin.ch/bfe/en/home/policy/energy-strategy-2050.html`.

[2] European Union. “Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity and repealing directive 2003/54/EC”. In: *Official Journal of the European Union* 211 (2009), 55.

[3] V.A. Evangelopoulos, P.S. Georgilakis, and N.D. Hatziargyriou. “Optimal operation of smart distribution networks: A review of models, methods and future research”. In: *Electric Power Systems Research* 140 (2016), 95.

[4] Y. Wang, Q. Chen, T. Hong, and C. Kang. “Review of smart meter data analytics: Applications, methodologies, and challenges”. In: *IEEE Transactions on Smart Grid* 10.3 (2018), 3125.

[5] R.J. Bessa. “Future trends for big data application in power systems”. In: *Big data application in power systems*. Elsevier, 2018, 223.

[6] R.J. Hyndman, X.A. Liu, and P. Pinson. “Visualizing big energy data: Solutions for this crucial component of data analysis”. In: *Power and Energy Magazine* 16.3 (2018), 18.

[7] Y. Sharon, A.M. Annaswamy, A.L. Motto, and A. Chakraborty. “Topology identification in distribution network with limited measurements”. In: *Innovative Smart Grid Technologies (ISGT)*. IEEE. 2012, 1.

[8] A. Primadianto and C.N. Lu. “A review on distribution system state estimation”. In: *IEEE Transactions on Power Systems* 32.5 (2016), 3875.

[9] C. Deb, F. Zhang, J. Yang, S.E. Lee, and K.W. Shah. “A review on time series forecasting techniques for building energy consumption”. In: *Renewable and Sustainable Energy Reviews* 74 (2017), 902.

[10] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour. *Energy forecasting: A review and outlook*. Tech. rep. Department of Operations Research and Business Intelligence, Wroclaw, 2020.

[11] S. Yang and C. Shen. “A review of electric load classification in smart grid environment”. In: *Renewable and Sustainable Energy Reviews* 24 (2013), 103.

[12] R. Gopinath, M. Kumar, C.P.C. Joshua, and K. Srinivas. “Energy management using non-intrusive load monitoring techniques - State-of-the-art and future research directions”. In: *Sustainable Cities and Society* (2020), 102411.

[13] O. Ma, N. Alkadi, P. Cappers, P. Denholm, J. Dudley, S. Goli, M. Hummon, S. Kiliccote, J. MacDonald, and N. Matson. “Demand response for ancillary services”. In: *IEEE Transactions on Smart Grid* 4.4 (2013), 1988.

[14] J.X. Chin, T.T. De Rubira, and G. Hug. “Privacy-protecting energy management unit through model-distribution predictive control”. In: *IEEE Transactions on Smart Grid* 8.6 (2017), 3084.

[15] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin. “Peer-to-peer and community-based markets: A comprehensive review”. In: *Renewable and Sustainable Energy Reviews* 104 (2019), 367.

[16] J. Bloomberg. *Digitization, Digitalization, And Digital Transformation: Confuse Them At Your Peril.* URL: `https://www.forbes.com/sites/jasonbloomberg/2018/04/29/digitization-digitaliza%7B%7Dtion-and-digital-transformation-confuse-them-at-your-peril`.

[17] Gartner Glossary. URL: `https://www.gartner.com/en/information-technology/glossary/digitalization`.

[18] K. Schwab. *The Fourth Industrial Revolution: what it means, how to respond.* URL: `https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond`.

[19] Rethink Electric. *What is a Smart Grid and Why are We Building it?* URL: `https://rethinkelectric.com/what-is-a-smart-grid-why-are-we-building-it`.

[20] Q. Sun, H. Li, Z. Ma, C. Wang, J. Campillo, Q. Zhang, F. Wallin, and J. Guo. “A comprehensive review of smart energy meters in intelligent energy networks”. In: *IEEE Internet of Things Journal* 3.4 (2015), 464.

[21] European Commission. *Smart grids and meters.* URL: `https://ec.europa.eu/energy/topics/markets-and-consumers/smart-grids-and-meters/overview_en`.

[22] R.R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar. "A survey on advanced metering infrastructure". In: *International Journal of Electrical Power & Energy Systems* 63 (2014), 473.

[23] L. O'Malley. *The Evolving Digital Utility: The convergence of energy and IT.* Tech. rep. MaRS Discovery District, 2014.

[24] R. Arghandeh and Y. Zhou. *Big data application in power systems.* Elsevier, 2017.

[25] B. Pinte, M. Quinlan, and K. Reinhard. "Low voltage micro-phasor measurement unit (μPMU)". In: *Power and Energy Conference at Illinois (PECI).* IEEE. 2015, 1.

[26] A. Angioni, G. Lipari, M. Pau, F. Ponci, and A. Monti. "A low cost PMU to monitor distribution grids". In: *International Workshop on Applied Measurements for Power Systems (AMPS).* IEEE. 2017, 1.

[27] A. Von Meier, D. Culler, A. McEachern, and R. Arghandeh. "Micro-synchrophasors for distribution systems". In: *Innovative Smart Grid Technologies (ISGT).* IEEE. 2014.

[28] A. Chauhan and S. Rajvanshi. "Non-technical losses in power system: A review". In: *International Conference on Power, Energy and Control (ICPEC).* IEEE. 2013, 558.

[29] S.C. Yip, K. Wong, W.P. Hew, M.T. Gan, Raphael C.W. Phan, and S.W. Tan. "Detection of energy theft and defective smart meters in smart grids using linear regression". In: *International Journal of Electrical Power & Energy Systems* 91 (2017), 230.

[30] M. Kezunovic, P. Pinson, Z. Obradovic, S. Grijalva, T. Hong, and R. Bessa. "Big data analytics for future electricity grids". In: *Electric Power Systems Research* 189 (2020), 106788.

[31] U.S Department of Energy. "Advanced metering infrastructure and customer systems - Results from the smart grid investment grant program". In: (2016).

[32] A. Ghosal and M. Conti. "Key management systems for smart grid advanced metering infrastructure: A survey". In: *IEEE Communications Surveys & Tutorials* 21.3 (2019), 2831.

[33] A.R. Hambley, N. Kumar, and A.R. Kulkarni. *Electrical engineering: principles and applications.* Pearson Prentice Hall Upper Saddle River, NJ, 2008.

[34] D.B. Avancini, J. Rodrigues, S. Martins, R. Rabêlo, J. Al-Muhtadi, and P. Solic. "Energy meters evolution in smart grids: A review". In: *Journal of Cleaner Production* 217 (2019), 702.

[35] G.A. Ajenikoko and A.A. Olaomi. "Hardware design of a smart meter". In: *International Journal of Engineering Research and Applications (IJERA)* 4.6 (2014), 115.

[36] A. Jain and H. Singabhattu. "Multi-communication technology based AMI for smart metering in India". In: *International Conference for Convergence in Technology (I2CT).* IEEE. 2019, 1.

[37] K. Aravind, V. Cecchi, and A. Mukherjee. "A survey of communications and networking technologies for energy management in buildings and home automation". In: *Journal of Computer Networks and Communications* 2012 (2012).

[38] G.M. Toschi, L.B. Campos, and C.E. Cugnasca. "Home automation networks: A survey". In: *Computer Standards & Interfaces* 50 (2017), 42.

[39] W. Yan, Z. Wang, H. Wang, W. Wang, J. Li, and X. Gui. "Survey on recent smart gateways for smart home: Systems, technologies, and challenges". In: *Transactions on Emerging Telecommunications Technologies* (2020).

[40] B. Zhou, W. Li, K.W. Chan, Y. Cao, Y. Kuang, X. Liu, and X. Wang. "Smart home energy management systems: Concept, configurations, and scheduling strategies". In: *Renewable and Sustainable Energy Reviews* 61 (2016), 30.

[41] S. Tanakornpintong, N. Tangsunantham, T. Sangsuwan, and C. Pirak. "Location optimization for data concentrator unit in IEEE 802.15. 4 smart grid networks". In: *International Symposium on Communications and Information Technologies (ISCIT).* IEEE. 2017, 1.

[42] W. Meng, R. Ma, and H.H. Chen. "Smart grid neighborhood area networks: a survey". In: *IEEE Network* 28.1 (2014), 24.

[43] M.F. Khan, A. Jain, V. Arunachalam, and A. Paventhan. "Communication technologies for smart metering infrastructure". In: *Students' Conference on Electrical, Electronics and Computer Science.* IEEE. 2014, 1.

[44] M. Kuzlu, M. Pipattanasomporn, and S. Rahman. “Communication network requirements for major smart grid applications in HAN, NAN and WAN”. In: *Computer Networks* 67 (2014), 74.

[45] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, M.S. Obaidat, and J. Rodrigues. “Fog computing for smart grid systems in the 5G environment: Challenges and solutions”. In: *IEEE Wireless Communications* 26.3 (2019), 47.

[46] H.G. Schroder Filho, J. Pissolato Filho, and V.L. Moreli. “The adequacy of LoRaWAN on smart grids: A comparison with RF mesh technology”. In: *International Smart Cities Conference (ISC2)*. IEEE. 2016, 1.

[47] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi. “Big Data management in smart grid: concepts, requirements and implementation”. In: *Journal of Big Data* 4.1 (2017), 1.

[48] Y. Jiang, C.C. Liu, M. Diedesch, E. Lee, and A.K. Srivastava. “Outage management of distribution systems incorporating information from smart meters”. In: *IEEE Transactions on power systems* 31.5 (2015), 4144.

[49] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk. “Smart meter data analytics for distribution network connectivity verification”. In: *IEEE Transactions on Smart Grid* 6.4 (2015), 1964.

[50] H. Sæle and O.S. Grande. “Demand response from household customers: experiences from a pilot study in Norway”. In: *IEEE Transactions on Smart Grid* 2.1 (2011), 102.

[51] *Swiss Competence Centers for Energy Research*. URL: `https://www.innosuisse.ch/inno/en/home/thematische-programme/foerderprogramm-energie.html`.

[52] *Stromversorgungsverordnung*. URL: `https://www.admin.ch/opc/de/classified-compilation/20071266/index.html`.

[53] Ecoplan. *Smart Metering Roll Out – Kosten und Nutzen*. Tech. rep. 2015.

[54] *Elektrizitätswerk der Stadt Zürich (EWZ)*. URL: `https://www.ewz.ch/de/ueber-ewz/newsroom/medienmitteilungen/Faech%7B%7Dendeckende_Einfuehrung_intelligenter_Stromzaehler_ab_2021.html`.

[55] *Industrielle Werke Basel (IWB)*. URL: `https://www.iwb.ch`.

[56] *Elektrizitätswerke des Kantons Zürich (EKZ).* URL: `https://www.ekz.ch/de/ueber-ekz/engagement/smart-meter.html`.

[57] European Commission's Joint Research Centre on Smart Electricity Systems and Interoperability. *Smart Metering deployment in the European Union.* URL: `https://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union`.

[58] European Commission. "Benchmarking smart metering deployment in the EU-27 with a focus on electricity". In: (2014).

[59] European Commission. "Benchmarking Smart Metering Deployment in the EU-28". In: (2020).

[60] M. Kochański, K. Korczak, and T. Skoczkowski. "Technology Innovation System Analysis of Electricity Smart Metering in the European Union". In: *Energies* 13.4 (2020), 916.

[61] O. Bularca, M. Florea, and A.M. Dumitrescu. "Smart metering deployment status across EU-28". In: *International Symposium on Fundamentals of Electrical Engineering (ISFEE).* IEEE. 2018, 1.

[62] C. Stagnaro. "Second-generation smart meter roll-out in Italy: A cost-benefit analysis". In: *Power Summit.* Eurelectric. 2019.

[63] Y. Huang, E. Grahn, C.J. Wallnerström, L. Jaakonantti, and T. Johansson. "Smart Meters In Sweden - Lessons Learned and New Regulations". In: *Current and Future Challenges to Energy Security* (2018), 177.

[64] Selectra. *Linky : fonctionnement, installation, obligation.* URL: `https://selectra.info/energie/guides/compteurs/linky/deploiement`.

[65] Department for Business, Energy and Industrial Strategy. *Smart Meter Statistics in Great Britain: Quarterly Report to end June 2020.* Tech. rep. Government of Great Britain, 2020.

[66] BBC News. *Smart meter rollout delayed for four years.* URL: `https://www.bbc.com/news/business-49721436`.

[67] Smart Energy International. *COVID-19 threatens GB smart meter rollout plans.* URL: `https://www.smart-energy.com/industry-sectors/smart-meters/covid-19-threatens-gb-smart-meter-rollout-plans`.

[68] Dinheiro Vivo. *EDP Distribuição já investiu 130 milhões em contadores inteligentes*. URL: `https://www.dinheirovivo.pt/economia/edp-distribuicao-ja-investiu-130-milhoes-em-contadores-inteligentes-12780879.html`.

[69] German Bundestag. "Gesetz zur Digitalisierung der Energiewende". In: *Bundesgesetzblatt* 1.43 (2016).

[70] Federal Ministry of Economics and Technology (BMWi). *Smart Meter: Intelligente Messsysteme für die Energiewende Einleitung*. URL: `https://www.bmwi.de/Redaktion/DE/Textsammlungen/Energie/smart-meter.html`.

[71] E.ON Energie. *Smart Meter Pflicht: Stromzähler werden intelligent*. URL: `https://www.eon.de/de/eonerleben/smart-meter-pflicht-in-deutschland.html`.

[72] ADD Grup. *Against all odds: PRE Czech Republic continues installation of ADDAX PRIME based smart metering solution*. URL: `https://addgrup.com/es/news/against-all-odds-pre-czech-republic-continues-installation-of-addax-prime-based-smart-metering-solution`.

[73] ADD Grup. *550 000 ADDAX meters to be deployed by EVN Bulgaria*. URL: `https://addgrup.com/es/news/550-000-addax-meters-to-be-deployed-by-evn-bulgaria`.

[74] IoT Analytics. *Smart Meter Market 2019: Global penetration reached 14% – North America, Europe ahead*. URL: `https://iot-analytics.com/smart-meter-market-2019-global-penetration-reached-14-percent`.

[75] U.S. Energy Information Administration. "Annual Electric Power Industry Report, Form EIA-861 detailed data files". In: (2020).

[76] Natural Resources Canada. "Smart Grid in Canada". In: (2019).

[77] Government of Canada. *Smart Grid Program*. URL: `https://www.nrcan.gc.ca/climate-change/green-infrastructure-programs/smart-grids/19793`.

[78] Berg Insight. "Smart Metering in North America and Asia-Pacific". In: *M2M Research Series* (2019).

[79] Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE). *Los beneficios de los medidores inteligentes*. URL: `https://portal.ute.com.uy/noticias/los-beneficios-de-los-medidores-inteligentes`.

[80] La Republica. *Medidores inteligentes beneficiarán a 285 mil josefinos abonados de servicio eléctrico*. URL: `https://www.larepublica.net/noticia/medidores-inteligentes-beneficiaran-a-285-mil-josefinos-abonados-de-servicio-electrico`.

[81] bnamericas. *When will Latin America definitively switch to smart meters?* URL: `https://www.bnamericas.com/en/features/when-will-latin-america-definitively-switch-to-smart-meters`.

[82] Agemcamp. *CPFL vai instalar medidores inteligentes*. URL: `http://www.agemcamp.sp.gov.br/en/cpfl-vai-instalar-medidores-inteligentes`.

[83] Y. Zhou and R. Arghandeh. "Moving toward sgile machine learning for data analytics in power systems". In: *Big Data Application in Power Systems*. Elsevier, 2018, 77.

[84] E. Mocanu, P.H. Nguyen, and M. Gibescu. "Deep learning for power system data analysis". In: *Big data application in power systems*. Elsevier, 2018, 125.

[85] P. McDaniel and S. McLaughlin. "Security and privacy challenges in the smart grid". In: *IEEE Security & Privacy* 7.3 (2009), 75.

[86] J. Prasad and R. Samikannu. "Overview, issues and prevention of energy theft in smart grids and virtual power plants in Indian context". In: *Energy Policy* 110 (2017), 365.

[87] J.B. Leite and J.R.S. Mantovani. "Detecting and locating non-technical losses in modern distribution networks". In: *IEEE Transactions on Smart Grid* 9.2 (2016), 1023.

[88] J.L. Viegas, P.R. Esteves, R. Melício, V.M.F. Mendes, and S.M. Vieira. "Solutions for detection of non-technical losses in the electricity grid: A review". In: *Renewable and Sustainable Energy Reviews* 80 (2017), 1256.

[89] T. Ahmad, H. Chen, J. Wang, and Y. Guo. "Review of various modeling techniques for the detection of electricity theft in smart grid environment". In: *Renewable and Sustainable Energy Reviews* 82 (2018), 2916.

[90] Y. Wang, Q. Chen, and C. Kang. "Electricity theft detection". In: *Smart Meter Data Analytics*. Springer, 2020, 79.

[91] J. Nagi, K.S. Yap, S.K. Tiong, S.K. Ahmed, and M. Mohamad. “Non-technical loss detection for metered customers in power utility using support vector machines”. In: *IEEE Transactions on Power Delivery* 25.2 (2009), 1162.

[92] P. Jokar, N. Arianpoo, and V.C.M. Leung. “Electricity theft detection in AMI using customers’ consumption patterns”. In: *IEEE Transactions on Smart Grid* 7.1 (2015), 216.

[93] V. Ford, A. Siraj, and W. Eberle. “Smart grid energy fraud detection using artificial neural networks”. In: *Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*. IEEE. 2014, 1.

[94] C. Cody, V. Ford, and A. Siraj. “Decision tree learning for fraud detection in consumer energy consumption”. In: *International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2015, 1175.

[95] M.M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito. “Detection of non-technical losses using smart meter data and supervised learning”. In: *IEEE Transactions on Smart Grid* 10.3 (2018), 2661.

[96] L.A.P. Júnior, C.C.O. Ramos, D. Rodrigues, D.R. Pereira, A.N. de Souza, K.A.P. da Costa, and J.P. Papa. “Unsupervised non-technical losses identification through optimum-path forest”. In: *Electric Power Systems Research* 140 (2016), 413.

[97] M. Zanetti, E. Jamhour, M. Pellenz, M. Penna, V. Zambenedetti, and I. Chueiri. “A tunable fraud detection system for advanced metering infrastructure using short-lived patterns”. In: *IEEE Transactions on Smart grid* 10.1 (2017), 830.

[98] X. Yang, P. Zhao, X. Zhang, J. Lin, and W. Yu. “Toward a Gaussian-mixture model-based detection scheme against data integrity attacks in the smart grid”. In: *Internet of Things Journal* 4.1 (2016), 147.

[99] C.H. Lo and N. Ansari. “CONSUMER: A novel hybrid intrusion detection system for distribution networks in smart grid”. In: *IEEE Transactions on Emerging Topics in Computing* 1.1 (2013), 33.

[100] C.L. Su, W.H. Lee, and C.K. Wen. “Electricity theft detection in low voltage networks with smart meters using state estimation”. In: *International Conference on Industrial Technology (ICIT)*. IEEE. 2016, 493.

[101] A.A. Cárdenas, S. Amin, G. Schwartz, R. Dong, and S. Sastry. “A game theory model for electricity theft detection and privacy-aware control in AMI systems”. In: *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2012, 1830.

[102] S. Amin, G.A. Schwartz, A.A. Cárdenas, and S.S. Sastry. “Game-theoretic models of electricity theft detection in smart utility networks: Providing new capabilities with advanced metering infrastructure”. In: *Control Systems Magazine* 35.1 (2015), 66.

[103] A. Sanchez and W. Rivera. “Big data analysis and visualization for the smart grid”. In: *International Congress on Big Data (BigData Congress)*. IEEE. 2017, 414.

[104] A. Jarrah Nezhad, T.K. Wijaya, M. Vasirani, and K. Aberer. “SmartD: Smart meter data analytics dashboard”. In: *Proceedings of the 5th international conference on Future energy systems*. 2014, 213.

[105] A.A. Munshi and A.M. Yasser. “Big data framework for analytics in smart grids”. In: *Electric Power Systems Research* 151 (2017), 369.

[106] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato. “Identification of power distribution network topology via voltage correlation analysis”. In: *Conference on Decision and Control*. IEEE. 2013, 1659.

[107] S. Park, D. Deka, and M. Chertkov. “Exact topology and parameter estimation in distribution grids with minimal observability”. In: *Power Systems Computation Conference (PSCC)*. IEEE. 2018, 1.

[108] K. Soumalas, G. Messinis, and N. Hatziargyriou. “A data driven approach to distribution network topology identification”. In: *PowerTech*. IEEE. 2017, 1.

[109] A. Guzmán, A. Argüello, J. Quirós-Tortós, and G. Valverde. “Processing and correction of secondary system models in geographic information systems”. In: *IEEE Transactions on Industrial Informatics* 15.6 (2018), 3482.

[110] G. Cavraro, V. Kekatos, and S. Veeramachaneni. “Voltage analytics for power distribution network topology verification”. In: *IEEE Transactions on Smart Grid* 10.1 (2017), 1058.

[111] W. Wang and N. Yu. “Parameter estimation in three-phase power distribution networks using smart meter data”. In: *International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE. 2020, 1.

[112] Y. Liao, Y. Weng, and R. Rajagopal. "Urban distribution grid topology reconstruction via Lasso". In: *Power and Energy Society General Meeting (PESGM)*. IEEE. 2016, 1.

[113] Z. Tian, W. Wu, and B. Zhang. "A mixed integer quadratic programming model for topology identification in distribution network". In: *IEEE Transactions on Power Systems* 31.1 (2015), 823.

[114] K. Dehghanpour, Z. Wang, J. Wang, Y. Yuan, and F. Bu. "A survey on state estimation techniques and challenges in smart distribution systems". In: *IEEE Transactions on Smart Grid* 10.2 (2018), 2312.

[115] C. Kuster, Y. Rezgui, and M. Mourshed. "Electrical load forecasting models: A critical systematic review". In: *Sustainable cities and society* 35 (2017), 257.

[116] H.S. Hippert, C.E. Pedreira, and R.C. Souza. "Neural networks for short-term load forecasting: A review and evaluation". In: *IEEE Transactions on Power Systems* 16.1 (2001), 44.

[117] R.H. Inman, H.T.C. Pedro, and C.F.M. Coimbra. "Solar forecasting methods for renewable energy integration". In: *Progress in energy and combustion science* 39.6 (2013), 535.

[118] D.W. Van der Meer, J. Widén, and J. Munkhammar. "Review on probabilistic forecasting of photovoltaic power production and electricity consumption". In: *Renewable and Sustainable Energy Reviews* 81 (2018), 1484.

[119] P. Pinson. "Wind energy: Forecasting challenges for its operational management". In: *Statistical Science* 28.4 (2013), 564.

[120] Y. Zhang, J. Wang, and X. Wang. "Review on probabilistic forecasting of wind power generation". In: *Renewable and Sustainable Energy Reviews* 32 (2014), 255.

[121] S.C. Chan, K.M. Tsui, H.C. Wu, Y. Hou, Y.C. Wu, and F.F. Wu. "Load/price forecasting and managing demand response for smart grids: Methodologies and challenges". In: *signal processing magazine* 29.5 (2012), 68.

[122] R. Weron. "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International journal of forecasting* 30.4 (2014), 1030.

[123] R. Sevlian and R. Rajagopal. "A scaling law for short term load forecasting on varying levels of aggregation". In: *International Journal of Electrical Power & Energy Systems* 98 (2018), 350.

[124] J.X. Chin, T. Zufferey, E. Shyti, and G. Hug. “Load Forecasting of Privacy-Aware Consumers”. In: *PowerTech*. IEEE. 2019, 1.

[125] G. Gross and F.D. Galiana. “Short-term load forecasting”. In: *Proceedings of the IEEE* 75.12 (1987), 1558.

[126] T. Hong, J. Wilson, and J. Xie. “Long term probabilistic load forecasting and normalization with hourly information”. In: *IEEE Transactions on Smart Grid* 5.1 (2013), 456.

[127] T. Gneiting and M. Katzfuss. “Probabilistic forecasting”. In: *Annual Review of Statistics and Its Application* 1 (2014), 125.

[128] T. Hong and S. Fan. “Probabilistic electric load forecasting: A tutorial review”. In: *International Journal of Forecasting* 32.3 (2016), 914.

[129] J.S. Armstrong. “Combining forecasts”. In: *Principles of forecasting*. Springer, 2001, 417.

[130] B. Liu, J. Nowotarski, T. Hong, and R. Weron. “Probabilistic load forecasting via quantile regression averaging on sister forecasts”. In: *IEEE Transactions on Smart Grid* 8.2 (2015), 730.

[131] Y. Wang, Q. Chen, M. Sun, C. Kang, and Q. Xia. “An ensemble forecasting method for the aggregated load with subprofiles”. In: *IEEE Transactions on Smart Grid* 9.4 (2018), 3906.

[132] R.J. Hyndman, R.A. Ahmed, G. Athanasopoulos, and H.L. Shang. “Optimal combination forecasts for hierarchical time series”. In: *Computational statistics & data analysis* 55.9 (2011), 2579.

[133] T. Hong, J. Xie, and J. Black. “Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting”. In: *International Journal of Forecasting* 35.4 (2019), 1389.

[134] T. Zufferey, A. Lepouze, and G. Hug. “Inadequacy of standard algorithms and metrics for short-term load forecasts in low-voltage grids”. In: *PowerTech*. IEEE. 2019, 1.

[135] T. Gneiting. “Making and evaluating point forecasts”. In: *Journal of the American Statistical Association* 106.494 (2011), 746.

[136] S. Haben, J. Ward, D.V. Greetham, C. Singleton, and P. Grindrod. “A new error measure for forecasts of household-level, high resolution electrical energy consumption”. In: *International Journal of Forecasting* 30.2 (2014), 246.

[137] Y. Wang, D. Gan, M. Sun, N. Zhang, Z. Lu, and C. Kang. “Probabilistic individual load forecasting using pinball loss guided LSTM”. In: *Applied Energy* 235 (2019), 10.

[138] S. Koch. “Demand response methods for ancillary services and renewable energy integration in electric power systems”. PhD thesis. ETH Zurich, 2012.

[139] R. D’hulst, W. Labeeuw, B. Beusen, S. Claessens, G. Deconinck, and K. Vanthournout. “Demand response flexibility and flexibility potential of residential smart appliances: Experiences from large pilot test in Belgium”. In: *Applied Energy* 155 (2015), 79.

[140] M.G. Vayá. “Optimizing the electricity demand of electric vehicles: creating value through flexibility”. PhD thesis. ETH Zurich, 2015.

[141] J. Chan, A. Liptsey-Rahe, R. Devenish, and J. Jones. *Behavioral demand response ranking.* US Patent App. 14/458,143. 2015.

[142] M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen. “Clustering of connection points and load modeling in distribution systems”. In: *IEEE Transactions on Power Systems* 28.2 (2012), 1255.

[143] C. Beckel, L. Sadamori, T. Staake, and S. Santini. “Revealing household characteristics from smart meter data”. In: *Energy* 78 (2014), 397.

[144] Y. Wang, Q. Chen, D. Gan, J. Yang, D.S. Kirschen, and C. Kang. “Deep learning-based socio-demographic information identification from smart meter data”. In: *IEEE Transactions on Smart Grid* 10.3 (2018), 2593.

[145] I. Abubakar, S.N. Khalid, M.W. Mustafa, H. Shareef, and M. Mustapha. “Application of load monitoring in appliances’ energy management–A review”. In: *Renewable and Sustainable Energy Reviews* 67 (2017), 235.

[146] A. Albert and R. Rajagopal. “Thermal profiling of residential energy use”. In: *IEEE Transactions on Power Systems* 30.2 (2014), 602.

[147] K. Kouzelis, Z.H. Tan, B. Bak-Jensen, J.R. Pillai, and E. Ritchie. “Estimation of residential heat pump consumption for flexibility market applications”. In: *IEEE Transactions on Smart Grid* 6.4 (2015), 1852.

[148] Y. Ji and R. Rajagopal. “Demand and flexibility of residential appliances: An empirical analysis”. In: *Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2017, 1020.

[149] N. Sadeghianpourhamami, T. Demeester, D.F. Benoit, M. Strobbe, and C. Develder. “Modeling and analysis of residential flexibility: Timing of white good usage”. In: *Applied energy* 179 (2016), 790.

[150] M.E.H. Dyson, S.D. Borgeson, M.D. Tabone, and D.S. Callaway. “Using smart meter data to estimate demand response potential, with application to solar energy integration”. In: *Energy Policy* 73 (2014), 607.

[151] V. Gómez, M. Chertkov, S. Backhaus, and H.J. Kappen. “Learning price-elasticity of smart consumers in power distribution systems”. In: *International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2012, 647.

[152] A. van Stiphout, J. Engels, D. Guldentops, and G. Deconinck. “Quantifying the flexibility of residential electricity demand in 2050: a bottom-up approach”. In: *PowerTech*. IEEE. 2015, 1.

[153] R. Li, Z. Wang, C. Gu, F. Li, and H. Wu. “A novel time-of-use tariff design based on Gaussian Mixture Model”. In: *Applied energy* 162 (2016), 1530.

[154] C. Feng, Y. Wang, K. Zheng, and Q. Chen. “Smart meter data-driven customizing price design for retailers”. In: *IEEE Transactions on Smart Grid* 11.3 (2019), 2043.

[155] S. Bowles. *Microeconomics: Behavior, institutions, and evolution.* Princeton University Press, 2009.

[156] S. Gyamfi, S. Krumdieck, and T. Urmee. “Residential peak electricity demand response—Highlights of some behavioural issues”. In: *Renewable and Sustainable Energy Reviews* 25 (2013), 71.

[157] J. Torriti, M.G. Hassan, and M. Leach. “Demand response experience in Europe: Policies, programmes and implementation”. In: *Energy* 35.4 (2010), 1575.

[158] H.C. Gils. “Assessment of the theoretical demand response potential in Europe”. In: *Energy* 67 (2014), 1.

[159] L. Söder, P.D. Lund, H. Koduvere, T.F. Bolkesjø, G.H. Rossebø, E. Rosenlund-Soysal, K. Skytte, J. Katz, and D. Blumberga. “A review of demand side flexibility potential in Northern Europe”. In: *Renewable and Sustainable Energy Reviews* 91 (2018), 654.

[160] P. Pinson and H. Madsen. “Benefits and challenges of electrical demand response: A critical review”. In: *Renewable and Sustainable Energy Reviews* 39 (2014), 686.

[161] S. Nolan and M. O'Malley. "Challenges and barriers to demand response deployment and evaluation". In: *Applied Energy* 152 (2015), 1.

[162] N. Good, K.A. Ellis, and P. Mancarella. "Review and classification of barriers and enablers of demand response in the smart grid". In: *Renewable and Sustainable Energy Reviews* 72 (2017), 57.

[163] R.N. Boisvert and B.F. Neenan. *Social welfare implications of demand response programs in competitive electricity markets*. Tech. rep. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2003.

[164] A. Nilsson, C.J. Bergstad, L. Thuvander, D. Andersson, K. Andersson, and P. Meiling. "Effects of continuous feedback on households' electricity consumption: Potentials and barriers". In: *Applied Energy* 122 (2014), 17.

[165] G. Thanos, M. Minou, T. Ganu, V. Arya, D. Chakraborty, J. van Deventer, and G.D. Stamoulis. "Evaluating demand response programs by means of key performance indicators". In: *International Conference on Communication Systems and Networks (COMSNETS)*. IEEE. 2013, 1.

[166] E. Cutter, C.K. Woo, F. Kahrl, and A. Taylor. "Maximizing the value of responsive load". In: *The Electricity Journal* 25.7 (2012), 6.

[167] European Commission. *Study on the effective integration of demand energy recourses for providing flexibility to the electricity system*. Tech. rep. 2014.

[168] C. Zhang, J. Wu, C. Long, and M. Cheng. "Review of existing peer-to-peer energy trading projects". In: *Energy Procedia* 105 (2017), 2563.

[169] Open Utility. "A glimpse into the future of Britain's energy economy". In: *White Paper* (2016).

[170] Vandebron. URL: `https://vandebron.nl/`.

[171] A. Wörner, A. Meeuw, L. Ableitner, F. Wortmann, S. Schopfer, and V. Tiefenbeck. "Trading solar energy within the neighborhood: Field implementation of a blockchain-based electricity market". In: *Energy Informatics* 2.1 (2019), 11.

[172] E. Mengelkamp, J. Gärttner, K. Rock, S. Kessler, L. Orsini, and C. Weinhardt. "Designing microgrid energy markets: A case study: The Brooklyn Microgrid". In: *Applied Energy* 210 (2018), 870.

[173] Stop Smart Meters! URL: `https://stopsmartmeters.org`.

[174] France Info. *L'article à lire pour comprendre la fronde contre le compteur électrique Linky.* URL: `https://www.francetvinfo.fr/economie/linky/l-article-a-lire-pour-comprendre-la-fronde-contre-le-compteur-electrique-linky_3149551.html`.

[175] American Cancer Society. URL: `https://www.cancer.org/cancer/cancer-causes/radiation-exposure/smart-meters.html`.

[176] France Info. *Loiret : la police met en cause un compteur Linky après l'incendie d'une maison.* URL: `https://www.francetvinfo.fr/economie/linky/la-police-met-en-cause-un-compteur-linky-apres-l-incendie-d-une-maison-dans-le-loiret_2985381.html`.

[177] The Sun. *Are smart meters a FIRE hazard? BBC Watchdog investigation finds poorly fitted meters may have started blazes.* URL: `https://www.thesun.co.uk/money/4112372/bbc-investigates-smart-meters-safety-house-fires`.

[178] F. Leferink, C. Keyer, and A. Melentjev. "Static energy meter errors caused by conducted electromagnetic interference". In: *electromagnetic compatibility magazine* 5.4 (2016), 49.

[179] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin. "Private memoirs of a smart meter". In: *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building.* 2010, 61.

[180] U. Greveler, P. Glösekötterz, B. Justusy, and D. Loehr. "Multimedia content identification through smart meter power usage profiles". In: *Proceedings of the International Conference on Information and Knowledge Engineering (IKE).* The Steering Committee of The World Congress in Computer Science, Computer Engineering, and Applied Computing. 2012, 1.

[181] A. Hansen, J. Staggs, and S. Shenoi. "Security analysis of an advanced metering infrastructure". In: *International Journal of Critical Infrastructure Protection* 18 (2017), 3.

[182] S. McLaughlin, D. Podkuiko, S. Miadzvezhanka, A. Delozier, and P. McDaniel. "Multi-vendor penetration testing in the advanced metering infrastructure". In: *Proceedings of the 26th Annual Computer Security Applications Conference.* 2010, 107.

[183] S. Tweneboah-Koduah, A.K. Tsetse, J. Azasoo, and B. Endicott-Popovsky. “Evaluation of cybersecurity threats on smart metering system”. In: *Information Technology-New Generations*. Springer, 2018, 199.

[184] PricewaterhouseCoopers. *Consumer intelligence series: Protect.me*. Tech. rep. 2017.

[185] M.M. Pour, A. Anzalchi, and A. Sarwat. “A review on cyber security issues and mitigation methods in smart grid systems”. In: *SoutheastCon 2017*. IEEE. 2017, 1.

[186] R. Hoenkamp, G.B. Huitema, and A.J.C. de Moor-van Vugt. “The neglected consumer: the case of the smart meter rollout in the Netherlands”. In: *Renewable Energy Law and Policy Review* (2011), 269.

[187] General Data Protection Regulation. “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46”. In: *Official Journal of the European Union (OJ)* 59.1-88 (2016), 294.

[188] G. Le Ray and P. Pinson. “The ethical smart grid: Enabling a fruitful and long-lasting relationship between utilities and customers”. In: *Energy Policy* 140 (2020), 111258.

[189] G. Eibl and D. Engel. “Influence of data granularity on smart meter privacy”. In: *IEEE Transactions on Smart Grid* 6.2 (2014), 930.

[190] K. Xing, C. Hu, J. Yu, X. Cheng, and F. Zhang. “Mutual privacy preserving k-means clustering in social participatory sensing”. In: *IEEE Transactions on Industrial Informatics* 13.4 (2017), 2066.

[191] S. Salehkalaibar, F. Aminifar, and M. Shahidehpour. “Hypothesis testing for privacy of smart meters with side information”. In: *IEEE Transactions on Smart Grid* 10.2 (2017), 2059.

[192] C. Gonçalves, R.J. Bessa, and P. Pinson. “A critical overview of privacy-preserving approaches for collaborative forecasting”. In: *International Journal of Forecasting* 37.1 (2020), 322.

[193] Swisspower Innovation. *Swiss Energy Startup Map 2020*. URL: `https://innovation.swisspower.ch/swiss-energy-startup-map-2020`.

[194] Hive Power. URL: `https://hivepower.tech`.

[195] Zaphiro. URL: `https://zaphiro.ch`.

[196] Bitlumens. URL: `https://www.bitlumens.com`.

[197] Clemap. URL: `https://www.clemap.ch`.

[198] Adaptricity. URL: `https://www.adaptricity.com`.

[199] S. Koch, F. Ferrucci, A. Ulbig, and M. Koller. “Time-series simulations and assessment of smart grid planning options of distribution grids”. In: *International Conference on Electricity Distribution (CIRED)*. 2015.

[200] A. Ulbig, S. Koch, and C. Antonakopoulos. “Towards more cost-effective PV connection request assessments via time-series-based grid simulation and analysis”. In: *CIRED - Open Access Proceedings Journal* 2017.1 (2017), 2560.

[201] D. Toffanin and A. Ulbig. “Taming uncertainty in distribution grid planning – A scenario-based methodology for the analysis of impact of electric vehicles”. In: (2019).

[202] N. Stocker, A. Ulbig, and D. Toffanin. “Results of the Sologrid pilot project – Decentralized load management to increase the efficiency of local energy communities”. In: (2018).

[203] depsys. URL: `https://www.depsys.com`.

[204] O. Alizadeh-Mousavi and J. Jaton. *Method for estimating the topology of an electric power network using metering data*. 2019.

[205] J. Jaton, G. Besson, M. De Vivo, M. Carpita, M. Paolone, K. Christakou, C. Mugnier, and O. Alizadeh-Mousavi. *Method of determining mutual voltage sensitivity coefficients between a plurality of measuring nodes of an electric power network*. US Patent App. 16/095,130. 2020.

[206] M. Bozorg, O. Alizader-Mousavi, S. Wasterlain, and M. Carpita. “Model-less/measurement-based computation of voltage sensitivities in unbalanced electrical distribution networks: experimental validation”. In: *European Conference on Power Electronics and Applications (EPE’19 ECCE Europe)*. IEEE. 2019.

[207] P. Moutis and O. Mousavi. “Digital twin of distribution power transformer for real-time monitoring of medium voltage from low voltage measurements”. In: *IEEE Transactions on Power Delivery* (2020).

[208] P.P. Moutis and O. Alizadeh-Mousavi. “A practical proposal for state estimation at balanced, radial distribution systems”. In: *Innovative Smart Grid Technologies Europe (ISGT Europe)*. IEEE. 2019, 1.

[209] Exnaton. URL: `https://www.exnaton.com`.

[210] L. Ableitner, A. Meeuw, S. Schopfer, V. Tiefenbeck, F. Wortmann, and A. Wörner. "Quartierstrom – Implementation of a real world prosumer centric local energy market in Walenstadt, Switzerland". In: *arXiv preprint arXiv:1905.07242* (2019).

[211] L. Ableitner, V. Tiefenbeck, A. Meeuw, A. Wörner, E. Fleisch, and F. Wortmann. "User behavior in a real-world peer-to-peer electricity market". In: *Applied Energy* 270 (2020), 115061.

[212] S. Barsali. *Benchmark systems for network integration of renewable and distributed energy resources.* 2014.

[213] Fraunhofer IWES and University of Kassel. *CIGRE networks in pandapower.* URL: `https://pandapower.readthedocs.io/en/v1.4.1/networks/cigre.html`.

[214] K. Rudion, A. Orths, Z.A. Styczynski, and K. Strunz. "Design of benchmark of medium voltage distribution network for investigation of DG integration". In: *Power Engineering Society General Meeting.* IEEE. 2006, 6.

[215] K.P. Schneider, B.A. Mather, B.C. Pal, C.W. Ten, G.J. Shirek, H. Zhu, J.C. Fuller, J.L.R. Pereira, L.F. Ochoa, and L.R. de Araujo. "Analytic considerations and design basis for the IEEE distribution test feeders". In: *IEEE Transactions on Power Systems* 33.3 (2017), 3181.

[216] F.E. Postigo Marcos, C. Mateo Domingo, T. Gomez San Roman, B. Palmintier, B.M. Hodge, V. Krishnan, F. de Cuadra García, and B. Mather. "A review of power distribution test feeders in the United States and the need for synthetic representative networks". In: *Energies* 10.11 (2017), 1896.

[217] N. Pflugradt and B. Platzer. "Behavior based load profile generator for domestic hot water and electricity use". In: *International Conference on Energy Storage (Innostock), Lleida, Spain.* 2012.

[218] N. Pflugradt, J. Teuscher, B. Platzer, and W. Schufft. "Analysing low-voltage grids using a behaviour based load profile generator". In: *International Conference on Renewable Energies and Power Quality.* Vol. 11. 2013, 5.

[219] Commission for Energy Regulation. *CER smart metering project.* URL: `https://www.ucd.ie/issda/data/commissionforenergy%20regulationcer`.

[220] Dataport. *Pecan Street.* URL: `https://www.pecanstreet.org/dataport`.

[221] A. Ulbig, T. Zufferey, O. Rodriguez Villalon, and S. Koch. *Optimized distribution grid operation by utilization of smart metering data.* URL: `https://www.aramis.admin.ch/Texte/?ProjectID=35398`.

[222] Neplan. URL: `https://www.neplan.ch`.

[223] *Compañía Nacional de Fuerza y Luz (CNFL).* URL: `https://www.cnfl.go.cr`.

[224] Electric Power Research Institute (EPRI). *openDSS.* URL: `https://www.epri.com/pages/sa/opendss`.

[225] J. Quirós Tortós. *Methodology for determining load profiles and residential electric consumption per type of device - Final report (available in Spanish).* Tech. rep. School of Electrical Engineering, University of Costa Rica, 2019.

[226] J. Chambers. *The R project for statistical computing.* URL: `https://www.r-project.org/`.

[227] RStudio. URL: `https://rstudio.com/`.

[228] M. Dowle. *data.table.* URL: `https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html`.

[229] M. Wallig. *doSNOW.* URL: `https://www.rdocumentation.org/packages/doSNOW`.

[230] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr.* URL: `https://dplyr.tidyverse.org/`.

[231] H. Wickham. *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York, 2016.

[232] L. Wilkinson. “The grammar of graphics”. In: *Handbook of Computational Statistics.* Springer, 2012, 375.

[233] G. Grolemund and H. Wickham. “Dates and times made easy with lubridate”. In: *Journal of Statistical Software* 40.3 (2011), 1.

[234] C. Sievert. *plotly.* URL: `https://plotly.com/r/`.

[235] H. Wickham. *reshape2.* URL: `https://www.rdocumentation.org/packages/reshape2`.

[236] Electric Power Research Institute (EPRI). *openDSS.* URL: `https://smartgrid.epri.com/SimulationTool.aspx`.

[237] MeteoSwiss. URL: `https://gate.meteoswiss.ch/idaweb`.

[238] J.W. Tukey. *Exploratory data analysis.* Vol. 2. Reading, Mass., 1977.

[239] A. Blázquez-García, A. Conde, U. Mori, and J.A. Lozano. “A review on outlier/anomaly detection in time series data”. In: *arXiv preprint arXiv:2002.04236* (2020).

[240] J.S. Chou and A.S. Telaga. “Real-time detection of anomalous power consumption”. In: *Renewable and Sustainable Energy Reviews* 33 (2014), 400.

[241] X. Liu and P.S. Nielsen. “Scalable prediction-based online anomaly detection for smart meter data”. In: *Information Systems* 77 (2018), 34.

[242] K. Hollingsworth, K. Rouse, J. Cho, A. Harris, M. Sartipi, S. Sozer, and B. Enevoldson. “Energy anomaly detection with forecasting and deep learning”. In: *International Conference on Big Data*. IEEE. 2018, 4921.

[243] J. Zhao, K. Liu, W. Wang, and Y. Liu. “Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry”. In: *Information Sciences* 259 (2014), 335.

[244] Y. Zhang, W. Chen, and J. Black. “Anomaly detection in premise energy consumption data”. In: *Power and Energy Society General Meeting*. IEEE. 2011, 1.

[245] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

[246] A.N. Baraldi and C.K. Enders. “An introduction to modern missing data analyses”. In: *Journal of school psychology* 48.1 (2010), 5.

[247] R. Hyndman. *Neural Network Time Series Forecasts (NNETAR)*. URL: `https://www.rdocumentation.org/packages/forecast/versions/8.13/topics/nnetar`.

[248] I. Pratama, A.E. Permanasari, I. Ardiyanto, and R. Indrayani. “A review of missing values handling methods on time-series data”. In: *International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE. 2016, 1.

[249] J. Ma, J.C.P. Cheng, F. Jiang, W. Chen, M. Wang, and C. Zhai. “A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data”. In: *Energy and Buildings* 216 (2020), 109941.

[250] F. Lobato, C. Sales, I. Araujo, V. Tadaiesky, L. Dias, L. Ramos, and A. Santana. “Multi-objective genetic algorithm for missing data imputation”. In: *Pattern Recognition Letters* 68 (2015), 126.

[251] Empa and Eawag. *NEST*. URL: `https://www.empa.ch/web/nest/overview`.

[252] *Swiss Federal Laboratories for Materials Science and Technology (Empa)*. URL: `https://www.empa.ch/web/empa`.

[253] E. Handschin, F.C. Schweppe, J. Kohlas, and A.A.F.A. Fiechter. “Bad data analysis for power system state estimation”. In: *IEEE Transactions on Power Apparatus and Systems* 94.2 (1975), 329.

[254] M.S. Uddin, A. Kuh, Y. Weng, and M. Ilić. “Online bad data detection using kernel density estimation”. In: *Power and Energy Society General Meeting*. IEEE. 2015, 1.

[255] D. Waeresch, R. Brandalik, W.H. Wellssow, J. Jordan, R. Bischler, and N. Schneider. “Bad data processing for low voltage state estimation systems based on smart meter data”. In: *CIRED - Open Access Proceedings Journal* (2016).

[256] M. Cramer, P. Goergens, and A. Schnettler. “Bad data detection and handling in distribution grid state estimation using artificial neural networks”. In: *PowerTech*. IEEE. 2015, 1.

[257] Y. Weng, R. Negi, and M.D. Ilić. “Historical data-driven state estimation for electric power systems”. In: *International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2013, 97.

[258] M. Picallo, A. Anta, B. De Schutter, and A. Panosyan. “A two-step distribution system state estimator with grid constraints and mixed measurements”. In: *Power Systems Computation Conference (PSCC)*. IEEE. 2018, 1.

[259] S.F. Wu, C.Y. Chang, and S.J. Lee. “Time series forecasting with missing values”. In: *International Conference on Industrial Networks and Intelligent Systems (INISCom)*. IEEE. 2015, 151.

[260] Y.G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier, and A. Aït-Bachir. “Period-aware content attention RNNs for time series forecasting with missing values”. In: *Neurocomputing* 312 (2018), 177.

[261] T. Zufferey, A. Ulbig, S. Koch, and G. Hug. “Unsupervised learning methods for power system data analysis”. In: *Big data application in power systems*. Elsevier, 2018, 107.

[262] L. Wen, K. Zhou, S. Yang, and L. Li. “Compression of smart meter big data: A survey”. In: *Renewable and Sustainable Energy Reviews* 91 (2018), 59.

[263] R. Mehra, N. Bhatt, F. Kazi, and N.M. Singh. “Analysis of PCA based compression and denoising of smart grid data under normal and fault conditions”. In: *International Conference on Electronics, Computing and Communication Technologies*. IEEE. 2013, 1.

[264] S. Das and P.S.N. Rao. “Principal component analysis based compression scheme for power system steady state operational data”. In: *ISGT2011-India*. IEEE. 2011, 95.

[265] Z. Shao, S.L. Yang, and F. Gao. “Density prediction and dimensionality reduction of mid-term electricity demand in China: A new semiparametric-based additive model”. In: *Energy conversion and management* 87 (2014), 439.

[266] F. McLoughlin, A. Duffy, and M. Conlon. “A clustering approach to domestic electricity load profile characterisation using smart metering data”. In: *Applied energy* 141 (2015), 190.

[267] S. Haben, C. Singleton, and P. Grindrod. “Analysis and clustering of residential customers energy behavioral demand using smart meter data”. In: *IEEE Transactions on Smart Grid* 7.1 (2015), 136.

[268] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst. “Phase identification of smart meters by clustering voltage measurements”. In: *Power Systems Computation Conference (PSCC)*. IEEE. 2018, 1.

[269] G. Chicco. “Overview and performance assessment of the clustering methods for electrical load pattern grouping”. In: *Energy* 42.1 (2012), 68.

[270] G.J. Tsekouras, N.D. Hatziargyriou, and E.N. Dialynas. “Two-stage pattern recognition of load curves for classification of electricity customers”. In: *IEEE Transactions on Power Systems* 22.3 (2007), 1120.

[271] J.T. Tou and R.C. Gonzalez. “Pattern recognition principles”. In: (1974).

[272] R.A. Fisher. *Iris data set.* URL: `https://archive.ics.uci.edu/ml/datasets/iris`.

[273] J.L. Viegas, S.M. Vieira, R. Melício, V.M.F. Mendes, and J.M.C. Sousa. “Classification of new electricity customers based on surveys and smart metering data”. In: *Energy* 107 (2016), 804.

[274] M. Azaza and F. Wallin. “Smart meter data clustering using consumption indicators: responsibility factor and consumption variability”. In: *Energy Procedia* 142 (2017), 2236.

[275] V. Agafonkin. *Leaflet - a JavaScript library for interactive maps.* URL: `https://leafletjs.com`.

[276] R. Silipo and P. Winters. “Big data, smart energy, and predictive analytics”. In: *Time Series Prediction of Smart Energy Data* 1 (2013), 37.

[277] J. Rousseau, T. Zufferey, M. Händel, and G. Hug. “Impact of time and spatial aggregation of smart meter data”. Semester Thesis - ETH Zürich. 2020.

[278] Navigant, Tractabel Impact, and Sweco. “Supporting country fiches accompanying the report “Benchmarking smart metering deployment in the EU-28””. In: (2020).

[279] A. Wright and S. Firth. “The nature of domestic electricity-loads and effects of time averaging on statistics and on-site generation calculations”. In: *Applied Energy* 84.4 (2007), 389.

[280] R. Granell, C.J. Axon, and D.C.H. Wallom. “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles”. In: *IEEE Transactions on Power Systems* 30.6 (2014), 3217.

[281] X. Han, D. Liu, J. Liu, and L. Kong. “Sensitivity analysis of acquisition granularity of photovoltaic output power to capacity configuration of energy storage systems”. In: *Applied Energy* 203 (2017), 794.

[282] R. Amaro e Silva and M.C. Brito. “Impact of network layout and time resolution on spatio-temporal solar forecasting”. In: *Solar Energy* 163 (2018), 329.

[283] M.R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac. “Smart meter data privacy: A survey”. In: *Communications Surveys & Tutorials* 19.4 (2017), 2820.

[284] L. Zhang, J. Zhang, and Y.H. Hu. “A privacy-preserving distributed smart metering temporal and spatial aggregation scheme”. In: *Access* 7 (2019), 28372.

[285] O.R.M. Boudia, S.M. Senouci, and M. Feham. “Elliptic curve-based secure multidimensional aggregation for smart grid communications”. In: *Sensors Journal* 17.23 (2017), 7750.

[286] T. Zufferey, A. Ulbig, S. Koch, and G. Hug. “Forecasting of smart meter time series based on neural networks”. In: *International workshop on data analytics for renewable energy integration.* Springer. 2016, 10.

[287] M. Pipattanasomporn, M. Kuzlu, S. Rahman, and Y. Teklu. “Load profiles of selected major household appliances and their demand response opportunities”. In: *IEEE Transactions on Smart Grid* 5.2 (2013), 742.

[288] T. Zufferey, D. Toffanin, D. Toprak, A. Ulbig, and G. Hug. “Generating stochastic residential load profiles from smart meter data for an optimal power matching at an aggregate level”. In: *Power Systems Computation Conference (PSCC)*. IEEE. 2018, 1.

[289] G. Tsagarakis, A.J. Collin, and A.E. Kiprakis. “Modelling the electrical loads of UK residential energy users”. In: *International Universities Power Engineering Conference (UPEC)*. IEEE. 2012, 1.

[290] A.J. Collin, G. Tsagarakis, A.E. Kiprakis, and S. McLaughlin. “Development of low-voltage load models for the residential load sector”. In: *IEEE Transactions on Power Systems* 29.5 (2014), 2180.

[291] K. McKenna and A. Keane. “Open and closed-loop residential load models for assessment of conservation voltage reduction”. In: *IEEE Transactions on Power Systems* 32.4 (2016), 2995.

[292] R. Singh, B.C. Pal, and R.A. Jabr. “Distribution system state estimation through Gaussian mixture model of the load as pseudo-measurement”. In: *IET Generation, transmission & distribution* 4.1 (2010), 50.

[293] A. Angioni, T. Schlösser, F. Ponci, and A. Monti. “Impact of pseudo-measurements from new power profiles on state estimation in low-voltage grids”. In: *IEEE Transactions on Instrumentation and Measurement* 65.1 (2015), 70.

[294] E. Manitsas, R. Singh, B.C. Pal, and G. Strbac. “Distribution system state estimation using an artificial neural network approach for pseudo measurement modeling”. In: *IEEE Transactions on power systems* 27.4 (2012), 1888.

[295] Y.R. Gahrooei, A. Khodabakhshian, and R.A. Hooshmand. “A new pseudo load profile determination approach in low voltage distribution networks”. In: *IEEE Transactions on Power Systems* 33.1 (2017), 463.

[296] W. Soares, J.C.S. de Souza, M.B. Do Coutto Filho, and A.A. Augusto. “Distribution system state estimation with real-time pseudo-measurements”. In: *Innovative Smart Grid Technologies Latin America (ISGT Latin America)*. IEEE. 2019, 1.

[297] Z. Cao, Y. Wang, C.C. Chu, and R. Gadh. “Robust pseudo-measurement modeling for three-phase distribution systems state estimation”. In: *Electric Power Systems Research* 180 (2020), 106138.

[298] Y. Wang, Q. Chen, and C. Kang. “Residential load data generation”. In: *Smart Meter Data Analytics*. Springer, 2020, 99.

[299] W.H. Kersting and W.H. Phillips. “Load allocation based upon automatic meter readings”. In: *IEEE/PES Transmission and Distribution Conference and Exposition* (2008), 1.

[300] D. Groß, P. Wiest, and K. Rudion. “Comparison of stochastic load profile modeling approaches for low voltage residential consumers”. In: *PowerTech, Manchester* (2017).

[301] W. Labeeuw and G. Deconinck. “Residential electrical load model based on mixture model clustering and Markov models”. In: *IEEE Transactions on Industrial Informatics* 9.3 (2013), 1561.

[302] F. McLoughlin, A. Duffy, and M. Conlon. “The generation of domestic electricity load profiles through markov chain modelling”. In: *Proceedings of 3rd International Scientific Conference on Energy and Climate Change* (2010).

[303] W. Labeeuw and G. Deconinck. “Customer sampling in a smart grid pilot”. In: *Power and Energy Society General Meeting* (2012).

[304] M. Uhrig, R. Mueller, and T. Leibfried. “Statistical consumer modelling based on smart meter measurement data”. In: *Proceedings of Probabilistic Methods Applied to Power Systems (PMAPS)* (2014).

[305] R. Singh, B.C. Pal, and R.A. Jabr. “Statistical representation of distribution system loads using Gaussian mixture model”. In: *IEEE Transactions on Power Systems* 25.1 (2010), 29.

[306] J. Bilmes. “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models”. In: *International Computer Science Institute* 4.510 (1998), 126.

[307] S. Agovski, T. Zufferey, N. Stocker, and G. Hug. “Synthesis of reactive power profiles for consumers in a distribution grid”. Semester Thesis - ETH Zürich. 2018.

[308] T. Zufferey and G. Hug. “Impact of data availability and pseudo-measurement synthesis on distribution system state estimation”. In: *IET Smart Grid* 5.1 (2021), 29.

[309] N. Karmarkar and R. M. Karp. “An efficient approximation scheme for the one-dimensional bin-packing problem”. In: *Annual Symposium on Foundations of Computer Science* (1982).

[310] J. Löfberg. *YALMIP*. URL: `https://yalmip.github.io/`.

[311] Gurobi Optimization. *Gurobi*. URL: `https://www.gurobi.com/`.

[312] P.A. Pegoraro, J. Tang, J. Liu, F. Ponci, A. Monti, and C. Muscas. “PMU and smart metering deployment for state estimation in active distribution grids”. In: *International Energy Conference and Exhibition (ENERGYCON)*. IEEE. 2012, 873.

[313] J. Liu, F. Ponci, A. Monti, C. Muscas, P.A. Pegoraro, and S. Sulis. “Optimal meter placement for robust measurement systems in active distribution grids”. In: *IEEE Transactions on Instrumentation and Measurement* 63.5 (2014), 1096.

[314] M.G. Damavandi, V. Krishnamurthy, and J.R. Martí. “Robust meter placement for state estimation in active distribution systems”. In: *IEEE Transactions on Smart Grid* 6.4 (2015), 1972.

[315] M. Armendariz, D. Babazadeh, L. Nordström, and M. Barchiesi. “A method to place meters in active low voltage distribution networks using BPSO algorithm”. In: *Power Systems Computation Conference (PSCC)*. IEEE. 2016, 1.

[316] T.C. Xygkis and G.N. Korres. “Optimized measurement allocation for power distribution systems using mixed integer SDP”. In: *IEEE Transactions on Instrumentation and measurement* 66.11 (2017), 2967.

[317] M. Humayun, J. Schoene, B. Poudel, B. Russell, G. Sun, J. Bui, A. Salazar, N. Badayos, M. Lak, and C.R. Clarke. “Quantifying distribution system state estimation accuracies achieved by adding telemetry and operational forecasting”. In: *SoutheastCon*. IEEE. 2019, 1.

[318] A. Abur and A. Gómez-Expósito. *Power system state estimation: theory and implementation*. CRC press, 2004.

[319] A. De la Villa Jaén, J.B. Martínez, A. Gómez-Expósito, and F.G. Vázquez. “Tuning of measurement weights in state estimation: Theoretical analysis and case study”. In: *IEEE Transactions on Power Systems* 33.4 (2017), 4583.

[320] *Matpower*. URL: `https://matpower.org/`.

[321] A.S. Masoum, P.S. Moses, M.A.S. Masoum, and A. Abu-Siada. “Impact of rooftop PV generation on distribution transformer and voltage profile of residential and commercial networks”. In: *Innovative Smart Grid Technologies (ISGT).* IEEE. 2012, 1.

[322] R. Scharrenberg, B. Vonk, and P.H. Nguyen. “EV stochastic modelling and its impacts on the Dutch distribution network”. In: *International Conference on Probabilistic Methods Applied to Power Systems (PMAPS).* IEEE. 2014, 1.

[323] B. Brinkmann and M. Negnevitsky. “A probabilistic approach to observability of distribution networks”. In: *IEEE Transactions on Power Systems* 32.2 (2016), 1169.

[324] R. Singh, B.C. Pal, R.A. Jabr, and R.B. Vinter. “Meter placement for distribution system state estimation: An ordinal optimization approach”. In: *IEEE Transactions on Power Systems* 26.4 (2011), 2328.

[325] S. Prasad and D.M.V. Kumar. “Trade-offs in PMU and IED deployment for active distribution state estimation using multi-objective evolutionary algorithm”. In: *IEEE Transactions on Instrumentation and Measurement* 67.6 (2018), 1298.

[326] R. Singh, B.C. Pal, and R.B. Vinter. “Measurement placement in distribution system state estimation”. In: *IEEE Transactions on Power Systems* 24.2 (2009), 668.

[327] A. Schrijver. *Combinatorial optimization: Polyhedra and efficiency.* Vol. 24. Springer Science & Business Media, 2003.

[328] G. Seifert, J. Gutbrod, M. Luther, and A. Kopken. “Reduction of measurement points in low-voltage grids with high PV-share”. In: *CIRED - Open Access Proceedings Journal* (2016).

[329] R. Bessa, G. Sampaio, V. Miranda, and J. Pereira. “Probabilistic low-voltage state estimation using analog-search techniques”. In: *Power Systems Computation Conference (PSCC).* IEEE. 2018, 1.

[330] F. Therrien, I. Kocar, and J. Jatskevich. “A unified distribution system state estimator using the concept of augmented matrices”. In: *IEEE Transactions on Power Systems* 28.3 (2013), 3390.

[331] P.M. De Oliveira-De Jesus and A.A.R. Quintana. “Distribution system state estimation model using a reduced quasi-symmetric impedance matrix”. In: *IEEE Transactions on Power Systems* 30.6 (2014), 2856.

[332] X. Zhou, Z. Liu, Y. Guo, C. Zhao, J. Huang, and L. Chen. "Gradient-based multi-area distribution system state estimation". In: *IEEE Transactions on Smart Grid* 11.6 (2020), 5325.

[333] U. Kuhar, M. Pantoš, G. Kosec, and A. Švigelj. "The impact of model and measurement uncertainties on a state estimation in three-phase distribution networks". In: *IEEE Transactions on Smart Grid* 10.3 (2018), 3301.

[334] P. Pinson. "Estimation of the uncertainty in wind power forecasting". PhD thesis. École Nationale Supérieure des Mines de Paris, 2006.

[335] T. Zufferey, G. Hug, and G. Valverde. "Disaggregation of cold appliance loads from smart meter data processing". In: *PES Transmission & Distribution Conference and Exhibition-Latin America (T&D LA)*. IEEE. 2020, 1.

[336] T. Zufferey, G. Hug, and G. Valverde. "Unsupervised disaggregation of water heater load from smart meter data processing". In: *Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MedPower)*. IEEE. 2020, 1.

[337] T. Ericson. "Direct load control of residential water heaters". In: *Energy Policy* 37.9 (2009), 3502.

[338] I.A. Sajjad, G. Chicco, M. Aziz, and A. Rasool. "Potential of residential demand flexibility-Italian scenario". In: *International Multi-Conference on Systems, Signals & Devices (SSD14)*. IEEE. 2014, 1.

[339] S. Koch, J.L. Mathieu, and D.S. Callaway. "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services". In: *Proc. PSCC*. Citeseer. 2011, 1.

[340] A. Kipping and E. Trømborg. "Modeling and disaggregating hourly electricity consumption in Norwegian dwellings based on smart meter data". In: *Energy and Buildings* 118 (2016), 350.

[341] T. Clarke, T. Slay, C. Eustis, and R.B. Bass. "Aggregation of residential water heaters for peak shifting and frequency response services". In: *Open Access Journal of Power and Energy* 7 (2019), 22.

[342] J. Ponoćko and J.V. Milanović. "Application of data analytics for advanced demand profiling of residential load using smart meter data". In: *PowerTech*. IEEE. 2017, 1.

[343] J. Ponoćko and J.V. Milanović. “Data requirements for a reliable demand decomposition in sparsely monitored power networks”. In: *Innovative Smart Grid Technologies Europe (ISGT Europe)*. IEEE. 2018, 1.

[344] A. Ridi, C. Gisler, and J. Hennebert. “A survey on intrusive load monitoring for appliance recognition”. In: *International Conference on Pattern Recognition*. IEEE. 2014, 3702.

[345] D.C. Mocanu, E. Mocanu, P.H. Nguyen, M. Gibescu, and A. Liotta. “Big IoT data mining for real-time energy disaggregation in buildings”. In: *International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2016, 3765.

[346] W. Kong, Z.Y. Dong, J. Ma, D.J. Hill, J. Zhao, and F. Luo. “An extensible approach for non-intrusive load disaggregation with smart meter data”. In: *IEEE Transactions on Smart Grid* 9.4 (2016), 3362.

[347] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han. “Unsupervised disaggregation of low frequency power measurements”. In: *Proceedings of the 2011 SIAM international conference on data mining*. SIAM. 2011, 747.

[348] N. Czarnek, K. Morton, L. Collins, R. Newell, and K. Bradbury. “Performance comparison framework for energy disaggregation systems”. In: *International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2015, 446.

[349] B. Huang, M. Knox, K. Bradbury, L.M. Collins, and R.G. Newell. “Non-intrusive load monitoring system performance over a range of low frequency sampling rates”. In: *International Conference on Renewable Energy Research and Applications (ICRERA)*. IEEE. 2017, 505.

[350] M. Stadler, W. Krause, M. Sonnenschein, and U. Vogel. “Modelling and evaluation of control schemes for enhancing load shift of electricity demand for cooling devices”. In: *Environmental Modelling & Software* 24.2 (2009), 285.

[351] M.A. Zehir and M. Bagriyanik. “Demand side management by controlling refrigerators and its effects on consumers”. In: *Energy Conversion and Management* 64 (2012), 238.

[352] G. Niro, D. Salles, M.V.P. Alcântara, and L.C.P. da Silva. “Large-scale control of domestic refrigerators for demand peak reduction in distribution systems”. In: *Electric power systems research* 100 (2013), 34.

[353] Samsung. *French door refrigerator - User manual.* Tech. rep. 2012.

[354] LG. *Demand response potential of residential appliances – Refrigerator.* Tech. rep. 2017.

[355] J.M. Belman-Flores, D. Pardo-Cely, M.A. Gómez-Martínez, I. Hernández-Pérez, D.A. Rodríguez-Valderrama, and Y. Heredia-Aricapa. "Thermal and energy evaluation of a domestic refrigerator under the influence of the thermal load". In: *Energies* 12.3 (2019), 400.

[356] B. Schaule, T. Zufferey, S. Koch, and G. Hug. "Disaggregation of smart meter measurement data into specific load components". Master Thesis - ETH Zürich. 2017.

[357] T. Zufferey, S. Renggli, and G. Hug. "Probabilistic state forecasting and optimal voltage control in distribution grids under uncertainty". In: *Electric Power Systems Research* 188 (2020), 106562.

[358] J. Xu, M. Yue, D. Katramatos, and S. Yoo. "Spatial-temporal load forecasting using AMI data". In: *International Conference on Smart Grid Communications (SmartGridComm).* IEEE. 2016, 612.

[359] B. Hayes, J. Gruber, and M. Prodanovic. "Short-term load forecasting at the local level using smart meter data". In: *PowerTech.* IEEE. 2015, 1.

[360] M. Ghofrani, M. Hassanzadeh, M. Etezadi-Amoli, and M.S. Fadali. "Smart meter based short-term load forecasting for residential customers". In: *North American Power Symposium (NAPS).* IEEE. 2011, 1.

[361] F.L. Quilumba, W.J. Lee, H. Huang, D.Y. Wang, and R.L. Szabados. "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities". In: *IEEE Transactions on Smart Grid* 6.2 (2014), 911.

[362] P. Lusis, K.R. Khalilpour, L. Andrew, and A. Liebman. "Short-term residential load forecasting: Impact of calendar effects and forecast granularity". In: *Applied Energy* 205 (2017), 654.

[363] A. Tascikaraoglu and B.M. Sanandaji. "Short-term residential electric load forecasting: A compressive spatio-temporal approach". In: *Energy and Buildings* 111 (2016), 380.

[364] L. Vallance, B. Charbonnier, N. Paul, S. Dubost, and P. Blanc. “Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric”. In: *Solar Energy* 150 (2017), 408.

[365] A. Elvers, M. Voß, and S. Albayrak. “Short-term probabilistic load forecasting at low aggregation levels using convolutional neural networks”. In: *PowerTech*. IEEE. 2019, 1.

[366] S. Karagiannopoulos, L. Roald, P. Aristidou, and G. Hug. “Operational planning of active distribution grids under uncertainty”. In: *IREP 2017, X Bulk Power Systems Dynamics and Control Symposium*. 2017.

[367] T. Soares, R.J. Bessa, P. Pinson, and H. Morais. “Active distribution grid management based on robust AC optimal power flow”. In: *IEEE Transactions on Smart Grid* 9.6 (2017), 6229.

[368] Y.P. Agalgaonkar, B.C. Pal, and R.A. Jabr. “Stochastic distribution system operation considering voltage regulation risks in the presence of PV generation”. In: *IEEE Transactions on Sustainable Energy* 6.4 (2015), 1315.

[369] D. Bertsimas, E. Litvinov, X.A. Sun, J. Zhao, and T. Zheng. “Adaptive robust optimization for the security constrained unit commitment problem”. In: *IEEE Transactions on Power Systems* 28.1 (2012), 52.

[370] L. Roald and G. Andersson. “Chance-constrained AC optimal power flow: Reformulations and efficient algorithms”. In: *IEEE Transactions on Power Systems* 33.3 (2017), 2906.

[371] S. Karagiannopoulos, P. Aristidou, and G. Hug. “Data-driven local control design for active distribution grids using off-line optimal power flow and machine learning techniques”. In: *IEEE Transactions on Smart Grid* 10.6 (2019), 6461.

[372] R.E. Edwards, J. New, and L.E. Parker. “Predicting future hourly residential electrical consumption: A machine learning case study”. In: *Energy and Buildings* 49 (2012), 591.

[373] S. Ryu, J. Noh, and H. Kim. “Deep neural network based demand side short term load forecasting”. In: *Energies* 10.1 (2017), 3.

[374] H. Shi, M. Xu, and R. Li. “Deep learning for household load forecasting—A novel pooling deep RNN”. In: *IEEE Transactions on Smart Grid* 9.5 (2017), 5271.

[375] R.J. Hyndman. *Forecasting with long seasonal periods*. 2010. URL: `https://robjhyndman.com/hyndsight/longseasonality`.

[376] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik. “Support vector regression machines”. In: *Advances in neural information processing systems* 9 (1997), 155.

[377] A.J. Smola and B. Schölkopf. “A tutorial on support vector regression”. In: *Statistics and computing* 14.3 (2004), 199.

[378] C.C. Chang and C.J. Lin. “LIBSVM: a library for support vector machines”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011), 1.

[379] B. Laperre, J. Amaya, and G. Lapenta. “Dynamic Time Warping as a New Evaluation for Dst Forecast with Machine Learning”. In: *arXiv preprint arXiv:2006.04667* (2020).

[380] E.H. Bristol. “Swinging door trending: Adaptive trend recording?” In: *ISA National Conference Proceedings*. 1990, 749.

[381] X. Wang, K. Smith-Miles, and R.J. Hyndman. “Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series”. In: *Neurocomputing* 72.10-12 (2009), 2581.

[382] D. Gan, Y. Wang, S. Yang, and C. Kang. “Embedding based quantile regression neural network for probabilistic load forecasting”. In: *Journal of Modern Power Systems and Clean Energy* 6.2 (2018), 244.

[383] R. Hecht-Nielsen. “Theory of the backpropagation neural network”. In: *Neural networks for perception*. Elsevier, 1992, 65.

[384] M.H. Hassoun. *Fundamentals of artificial neural networks*. MIT press, 1995.

[385] UVEK. *Wie viel Strom oder Wärme kann mein Dach produzieren?* URL: `www.uvek-gis.admin.ch/BFE/sonnendach`.

[386] Energie Schweiz. *Solarrechner*. URL: `www.energieschweiz.ch/page/de-ch/solarrechner`.

[387] J.D. Cross and R. Hartshorn. “My Electric Avenue: Integrating electric vehicles into the electrical networks”. In: (2016).

[388] Bosch Automotive Service Solutions. *Electric vehicle solutions*. URL: `www.boschevsolutions.com/charging-stations`.

[389] F. Ferrando, T. Zufferey, and G. Hug. “Potential impact of electric vehicles on the swiss energy system”. Master Thesis - ETH Zürich. 2019.

[390] J. Stiasny, T. Zufferey, G. Pareschi, D. Toffanin, G. Hug, and K. Boulouchos. “Sensitivity analysis of EV impact on distribution grids based on Monte-Carlo simulations”. Master Thesis - ETH Zürich. 2019.

[391] C. Wan, Z. Xu, Y. Wang, Z.Y. Dong, and K.P. Wong. “A hybrid approach for probabilistic forecasting of electricity price”. In: *IEEE Transactions on Smart Grid* 5.1 (2013), 463.

[392] S. Karagiannopoulos, P. Aristidou, A. Ulbig, S. Koch, and G. Hug. “Optimal planning of distribution grids considering active power curtailment and reactive power control”. In: *Power and Energy Society General Meeting (PESGM)*. IEEE. 2016, 1.

[393] G. Koeppel and D. Reichelt. “Power Market I”. In: *Lecture, ETH Zurich* (2020).

[394] European Energy Spot Market. *EPEX SPOT*. URL: `www.epexspot.com`.

[395] Statistisches Amt des Kantons Basel-Stadt. *Haushalte im Kanton Basel-Stadt*. URL: `www.statistik.bs.ch/zahlen/tabellen/1-bevoelkerung/haushalte.html`.

CURRICULUM VITAE

PERSONAL DATA

Name	Thierry Zufferey
Date of Birth	June 27, 1991
Place of Birth	Sierre (VS), Switzerland
Citizen of	Noble-Contrée (VS), Switzerland

EDUCATION

02/2014 – 09/2015	MSc in Electrical Engineering and Information Technology ETH Zurich, Switzerland *Major*: Energy systems and power electronics
09/2010 – 08/2013	BSc in Electrical Engineering and Information Technology ETH Zurich, Switzerland *Major*: Electric energy systems and mechatronics
08/2005 – 06/2010	High School Diploma Lycée-Collège cantonal de la Planta, Sion, Switzerland *Major*: Physics and applied mathematics

EMPLOYMENT

05/2016 – 06/2021	Research assistant at ETH Zurich, Switzerland Power Systems Laboratory *Advisor*: Prof. Dr. Gabriela Hug
09/2019 – 02/2020	Visiting researcher at University of Costa Rica, San José, Costa Rica Department of Power Systems and Electrical Machines *Advisor*: Prof. Dr. Gustavo Valverde

09/2015 – 12/2018	Energy data analyst at Adaptricity, Zurich, Switzerland Project Engineering Team *Advisor*: Dr. Andreas Ulbig
02/2015 – 05/2015	Experiment assistant at ETH Zurich, Switzerland Automatic Control Laboratory (IFA) *Advisor*: Prof. Dr. Roy Smith
09/2013 – 11/2013	Internship at Electricity Generating Authority of Thailand (EGAT), Bangkok, Thailand Hydro Electrical Department *Advisor*: Mr. Chanin Prueksapitak