TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Antti Mutanen

Improving Electricity Distribution System State Estimation with AMR-Based Load Profiles

Julkaisu 1529 • Publication 1529

Tampere 2018

Tampereen teknillinen yliopisto. Julkaisu 1529
Tampere University of Technology. Publication 1529

Antti Mutanen

Improving Electricity Distribution System State Estimation with AMR-Based Load Profiles

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 23rd of February 2018, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2018

Doctoral candidate: Antti Mutanen
Laboratory of Electrical Energy Engineering
Faculty of Computing and Electrical Engineering
Tampere University of Technology
Finland

Supervisors: Pertti Järventausta, Professor
Laboratory of Electrical Energy Engineering
Faculty of Computing and Electrical Engineering
Tampere University of Technology
Finland

Sami Repo, Professor
Laboratory of Electrical Energy Engineering
Faculty of Computing and Electrical Engineering
Tampere University of Technology
Finland

Pre-examiners: Gareth Taylor, Professor
Brunel Institute for Power Systems
Brunel University
United Kingdom

Lars Nordström, Professor
Department of Electric Power and Energy Systems
Royal Institute of Technology
Sweden

Opponents: Gareth Taylor, Professor
Brunel Institute for Power Systems
Brunel University
United Kingdom

Matti Lehtonen, Professor
Department of Electrical Engineering and Automation
Aalto University
Finland

ISBN 978-952-15-4094-3 (printed)
ISBN 978-952-15-4105-6 (PDF)
ISSN 1459-2045

Abstract

The ongoing battle against global warming is rapidly increasing the amount of renewable power generation, and smart solutions are needed to integrate these new generation units into the existing distribution systems. Smart grids answer this call by introducing intelligent ways of controlling the network and active resources connected to it. However, before the network can be controlled, the automation system must know what the node voltages and line currents defining the network state are.

Distribution system state estimation (DSSE) is needed to find the most likely state of the network when the number and accuracy of measurements are limited. Typically, two types of measurements are used in DSSE: real-time measurements and pseudo-measurements. In recent years, finding cost-efficient ways to improve the DSSE accuracy has been a popular subject in the literature. While others have focused on optimizing the type, amount and location of real-time measurements, the main hypothesis of this thesis is that it is possible to enhance the DSSE accuracy by using interval measurements collected with automatic meter reading (AMR) to improve the load profiles used as pseudo-measurements.

The work done in this thesis can be divided into three stages. In the first stage, methods for creating new AMR-based load profiles are studied. AMR measurements from thousands of customers are used to test and compare the different options for improving the load profiling accuracy. Different clustering algorithms are tested and a novel two-stage clustering method for load profiling is developed. In the second stage, a DSSE algorithm suited for smart grid environment is developed. Simulations and real-life demonstrations are conducted to verify the accuracy and applicability of the developed state estimator. In the third and final stage, the AMR-based load profiling and DSSE are combined. Matlab simulations with real AMR data and a real distribution network model are made and the developed load profiles are compared with other commonly used pseudo-measurements.

The results indicate that clustering is an efficient way to improve the load profiling accuracy. With the help of clustering, both the customer classification and customer class load profiles can be updated simultaneously. Several of the tested clustering algorithms were suited for clustering electricity customers, but the best results were achieved with a modified *k*-means algorithm. Results from the third stage simulations supported the main hypothesis that the new AMR-based load profiles improve the DSSE accuracy.

The results presented in this thesis should motivate distribution system operators and other actors in the field of electricity distribution to utilize AMR data and clustering algorithms in load profiling. It improves not only the DSSE accuracy but also many other functions that rely on load flow calculation and need accurate load estimates or forecasts.

Preface

The work presented in this thesis has been carried out during the years 2008–2017 in the Laboratory of Electrical Energy Engineering of Tampere University of Technology. This thesis has been supervised by Professor Pertti Järventausta and Professor Sami Repo to whom I would like to express my deepest gratitude. Without their hard work to set up new projects and to secure funding this thesis work would not have been possible.

The projects I have been working in, have been funded by the Finnish Funding Agency for Technology and Innovation (Tekes), European Union, Academy of Finland, Electricity Research Pool (ST-pooli) and countless companies. The support of these organizations is gratefully acknowledged. Financial support in the form of personal grants from the Fortum Foundation, KAUTE Foundation, Ulla Tuominen Foundation, Walter Ahlström Foundation, and TUT Foundation is also greatly appreciated.

During the years, I co-operated with several universities and research institutes and would like to thank all my collaborators in the Technical Research Centre of Finland, University of Eastern Finland, Lappeenranta University of Technology, University of Strathclyde, Charles III University of Madrid, RWTH Aachen, and Danish Energy Association. From the industrial partners, I would like to thank especially Elenia, Koillis-Satakunnan Sähkö, Satapirkan Sähkö, Unareti, Unión Fenosa Distribución, Østkraft, and ABB Distribution Automation unit in Hermia for enabling the demonstrations and providing measurement and network data essential to my thesis.

It goes without saying that I am deeply thankful to all my co-authors. In addition, I would like to thank Antti Rautiainen for his encouragement and comments to my manuscript, Hannu Reponen and Shengye Lu for helping me with SQL databases, and Ontrei Raipala for his help with the RTDS. Big thanks also to all those colleagues who made the lunch- and coffee breaks relaxing, played badminton, and had fun on conference and seminar trips.

Not forgetting the importance of a balanced civilian life, I would like to thank my family, friends, and girlfriend for their love and support.

Finally, I would like to thank the cold and rainy Finnish summer for motivating me to stay indoors and finalize this thesis.

Tampere, January 2018

Antti Mutanen

Table of contents

Abstract i
Preface iii
Table of contents v
List of publications vii
List of abbreviations ix
List of symbols xi
1 Introduction 1
1.1 Smart grid control and its challenges 1
1.2 Smart meter rollout and the ensuing opportunities 3
1.3 The evolution and scope of the thesis 3
1.4 Main contributions 6
1.5 Publications and author's contribution 7
2 Background to load profiling 9
2.1 Electricity meter reading 9
2.1.1 Automatic meter reading 9
2.1.2 Advanced metering 10
2.2 History of load profiling 10
2.2.1 History and present state of load profiling in Finland 12
2.2.2 History and present state of load profiling in some other countries 13
2.3 Load profiles in distribution network calculation 15
2.3.1 Distribution network calculation in Finland 16
2.3.2 Defects in the existing load profiles 18
3 Methods for improving load profiling 21
3.1 Load temperature dependency calculation 21
3.1.1 Calculation of temperature dependency parameters 23
3.2 Load profile updating 24
3.3 Customer reclassification 25
3.4 Individual load profiles 26
3.4.1 Comparison with other load profiling methods 28
3.4.2 Improvements to the type weeks 28
3.5 Geographically bounded load profiles 29

3.6 Customer behavior change detection and other possible improvements 30
4 Clustering of electricity customers 33
4.1 Clustering algorithms 33
4.1.1 Partitioning methods 34
4.1.2 Hierarchical methods 36
4.1.3 Density-based methods 37
4.1.4 Grid-based methods 38
4.2 Comparison of clustering methods 38
4.3 Dimension reduction 39
4.4 Selecting the optimal number of clusters 41
4.4.1 Cluster validity indices 41
4.4.2 Knee point detection 44
4.5 The developed load profiling procedure 46
4.5.1 Load profile updating 48
4.5.2 Two-stage clustering 48
4.5.3 Sensitivity to initialization and clustering method 50
5 Distribution system state estimation 55
5.1 Literature review 55
5.1.1 Weighted least squares estimation 55
5.1.2 Other DSSE methods 58
5.2 The choice of state estimator 60
5.3 The developed state estimator 61
5.3.1 State estimate uncertainties 62
5.3.2 Estimation of weakly meshed networks 62
5.3.3 Algorithm implementation 64
5.3.4 Decentralized DSSE 68
5.3.5 State forecasting 70
5.4 Review and discussion on the achieved results 71
6 Conclusions and future research 75
6.1 Future research topics 77
References 79

List of publications

This thesis is based on the following original publications, which are referred to in the text as [P1]–[P9].

[P1] **A. Mutanen**, S. Repo, and P. Järventausta, "AMR in distribution network state estimation," presented at the 8th Nordic Electricity Distribution and Asset Management Conference (NORDAC), Bergen, Norway, Sept. 8–9, 2008.

[P2] **A. Mutanen**, A. Koto, A. Kulmala, and P. Järventausta, "Development and testing of a branch current based distribution system state estimator," presented at the 46th International Universities' Power Engineering Conference (UPEC). Soest, Germany, Sep. 5–8, 2011.

[P3] **A. Mutanen**, S. Repo, P. Järventausta, A. Löf, and D.D. Giustina, "Testing low voltage network state estimation in RTDS environment," presented at the 4th IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Copenhagen, Denmark, Oct. 6–9, 2013.

[P4] **A. Mutanen**, S. Repo, and P. Järventausta, "Customer classification and load profiling based on AMR measurements," presented at the 21st International Conference and Exhibition on Electricity Distribution (CIRED). Frankfurt, Germany, June 6–9, 2011.

[P5] **A. Mutanen**, M. Ruska, S. Repo, and P. Järventausta, "Customer classification and load profiling method for distribution systems," IEEE Transactions on Power Delivery, vol. 26, no. 3, July 2011.

[P6] **A. Mutanen**, P. Järventausta, M. Kärenlampi, and P. Juuti, "Improving distribution network analysis with new AMR-based load profiles," presented at the 22nd International Conference and Exhibition on Electricity Distribution (CIRED), Stockholm, Sweden, June 10–13, 2013.

[P7] B. Stephen, **A. Mutanen**, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," IEEE Transactions on Power Delivery, vol. 29, no. 1, Feb. 2014.

[P8] P. Koponen, **A. Mutanen**, and H. Niska, "Assessment of some methods for short-term load forecasting," presented at the 5th IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Istanbul, Turkey, Oct. 12–15, 2014.

[P9] **A. Mutanen**, P. Järventausta, and S. Repo, "Smart meter data -based load profiles and their effect on distribution system state estimation accuracy," submitted to International Review of Electrical Engineering.

List of abbreviations

ABB	ASEA Brown Boveri
AIC	Akaike Information Criterion
AMI	Advanced Metering Infrastructure
AMM	Advanced Metering Management
AMR	Automatic Meter Reading
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CDI	Clustering Dispersion Indicator
CH	Calinski–Harabasz criterion
CIS	Customer Information System
CLIQUE	Clustering in Quest
COP	Coefficient of Performance
CURE	Clustering Using Representatives
CVC	Coordinated Voltage Control
DBI	Davies–Bouldin Index
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DDA	Descriptive Discriminant Analysis
DENCLUE	Density Based Clustering
DG	Distributed Generation
DMS	Distribution Management System
DR	Demand Response
DSO	Distribution System Operator
DSSE	Distribution System State Estimation
EM	Expectation-Maximization
EU	European Union
GMM	Gaussian Mixture Model
ISODATA	Iterative Self-Organizing Data Analysis Techniques
LV	Low Voltage
LVSE	Low Voltage Network State Estimation
MAFIA	Merging of Adaptive Intervals Approach to Spatial Data Mining
MANOVA	Multivariate Analysis of Variance
MAPE	Mean Absolute Percentage Error
MASE	Multi-Area State Estimation

MDMS	Meter Data Management System
MFA	Mixtures of Factor Analyzers
MIA	Mean Index Adequacy
MV	Medium Voltage
MVSE	Medium Voltage Network State Estimation
NBI	Normal-Boundary Intersection
NIS	Network Information System
NN	Neural Network
OPTICS	Ordering Points to Identify the Clustering Structure
PCA	Principal Component Analysis
PDF	Probability Density Function
PMU	Phasor Measurement Unit
PSAU	Primary Substation Automation Unit
PSO	Particle Swarm Optimization
PV	Photovoltaic
RLP	Representative Load Pattern
RTDS	Real-Time Digital Simulator
SCADA	Supervisory Control and Data Acquisition
SLY	Association of Finnish Electric Utilities (in Finnish: Sähkölaitosyhdistys)
SMI	Smart Metering Infrastructure
SOM	Self-Organizing Map
SSAU	Secondary Substation Automation Unit
SSE	Sum of Squared Errors
STING	Statistical Information Grid
TDP	Typical Daily Profile
U.K.	United Kingdom
WLS	Weighted Least Squares

List of symbols

$\boldsymbol{a}$	A vector of temperature dependency parameters
α_j	Line j current phasor angle
$\boldsymbol{B}$	Matrix of factor loadings
β_j	Line j impedance phasor angle
C_i	Cluster i
$\boldsymbol{c_i}$	Cluster i centroid
c_r	Constraint from the real part of Kirchhoff's voltage law
c_x	Constraint from the imaginary part of Kirchhoff's voltage law
$\boldsymbol{D}$	Diagonal matrix
d	Number of subspace dimensions
E	Yearly energy
E[X]	Expected value of a random variable X
Eps	Neighborhood radius
ε	Convergence threshold
$f(\boldsymbol{x};\boldsymbol{\theta})$	Probability distribution of an observation $\boldsymbol{x}$ assuming Gaussian mixture parameters $\boldsymbol{\theta}$
$\boldsymbol{G}$	Gain matrix
$\boldsymbol{H}$	Jacobian matrix
$\boldsymbol{h(y)}$	Measurement function vector
$h_i(\boldsymbol{y})$	Measured variable i as a function of state variables $\boldsymbol{y}$
$\bar{I}$	Current phasor
$\bar{I}_j$	Current phasor on line j
$\lvert\bar{I}_j\rvert$	Current magnitude on line j
$\bar{I}_{XY}$	Current phasor on a line between nodes X and Y
J_c	Objective function to be minimized in fuzzy c-means clustering
J_k	Objective function to be minimized in k-means clustering
$J(\boldsymbol{y})$	Objective function to be minimized in weighted least squares estimation
$J_i(\boldsymbol{y})$	Value of $J(\boldsymbol{y})$ after the ith iteration round
$\Delta J(\boldsymbol{y})$	Gradient of $J(\boldsymbol{y})$
$\boldsymbol{K}$	Jacobian of $\boldsymbol{o(y)}$
K_f	Curvature of the function f
k	Number of clusters
Λ	A set of branches forming a loop
λ_j	Loop direction coefficient for branch j

$MinPts$	Density treshold
m	Blending parameter in fuzzy *c*-means
$minD$	Minimum distance from the nearest cluster centroid
μ_i	Mean of the component i
$N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Likelihood value for observation $\boldsymbol{x}$ assuming it is Gaussian distributed with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
N_m	Number of measurements
n	Sample size
$\boldsymbol{o(y)}$	A vector of network states as a function of $\boldsymbol{y}$
P	Active power
P_{ag}	Aggregated active power
ΔP	Outdoor temperature dependent part of the load
P_h	Active power loss
P_p	Active power with excess probability p
p	Excess probability
q	Number of dimensions in the data
$\boldsymbol{\theta}$	A set of Gaussian mixture parameters
$\boldsymbol{R}$	Covariance matrix
R	Pearson product-moment correlation coefficient
r	Reliability level
r_{max}^N	Largest normalized residual
$\bar{S}$	Complex power
$\bar{S}_h$	Complex power loss
s	Sample estimate of the population standard deviation (sample standard deviation)
s_{ag}	Aggregated sample estimate of the population standard deviation
$\boldsymbol{\Sigma}_i$	Covariance matrix of the *i*th component
σ_i^2	Variance of measurement i
T	Temperature
T_{24}	Average temperature from the previous 24 hours
t	Time
τ	Critical value of the Student's t-distribution
$\bar{V}$	Voltage phasor
$\bar{V}_X$	Voltage phasor on node X
$\boldsymbol{W}$	Weight matrix
w_i	Weight of the *i*th component
$\bar{X}$	Sample mean

$\boldsymbol{x}$	A vector of data points
$\boldsymbol{y}$	State variable vector
$\Delta\boldsymbol{y}$	Linearized approximation of the state vector change
$\bar{Z}_j$	Complex impedance on line j
$\bar{Z}_{XY}$	Complex impedance on a line between nodes X and Y
$\boldsymbol{z}$	Measurement vector
z	Z-score
z_i	Value of measurement i
z_p	Z-score corresponding to excess probability p

1 Introduction

Global warming is challenging humanity to co-operate and take actions to reduce greenhouse gas emissions. The fact that 191 independent states have signed and 87 have ratified the Paris agreement shows that there is a worldwide consensus that ambitious efforts are needed to limit global warming and its adverse effects (UNFCCC 2016).

In the spirit of the Paris agreement, the European Union (EU) has drawn up a new 2030 climate and energy framework which sets three key targets: at least a 40 % cut in greenhouse gas emissions (from 1990 levels), at least a 27 % share for renewable energy, and at least a 27 % improvement in energy efficiency (SN 79/14). The previous 2020 climate and energy package and its *20-20-20 targets* (406/2009/EC; 2009/28/EC) already caused a lot of movement in the energy sector. To meet these new targets, the share of renewable energy sources in electricity production needs to be substantially increased.

The renewable energy sources are usually distributed over large geographical areas and this often leads to many relatively small production units which are connected to distribution networks. If the distributed generation (DG) is based on the wind or direct usage of solar irradiation, it is also highly intermittent. These properties cause problems as the existing distribution networks have not been designed to accommodate large amounts of DG and the power systems have limited ability to balance the demand and varying electricity production.

The number one solution for the above-mentioned problems is the so-called *smart grid*. This much-hyped concept has many forms and countless different definitions. One of the most extensive and quoted definition has been given by the European Regulators Group for Electricity & Gas:

> *A smart grid is an electricity network that can cost efficiently integrate the behaviour and actions of all users connected to it - generators, consumers and those that do both - in order to ensure economically efficient, sustainable power system with low losses and high levels of quality and security of supply and safety* (ERGEG 2009, pp. 18–19).

When compared with the preceding definitions, this emphasizes cost-efficiency instead of more obscure "*intelligence*". In a truly smart grid, the use of modern technology and intelligence is a mean to achieve the desired targets, not an end in itself. While many of the challenges associated with the increasing DG installations could be solved with traditional network reinforcements (thicker cables, bigger transformers etc.), it is often more economical to use distribution network automation and functionalities such as coordinated voltage control, automatic network reconfiguration, demand response and production curtailment (Schiavo et al. 2015; Karali et al. 2015).

1.1 Smart grid control and its challenges

With automation, the distribution network utilization rates can be increased and the networks can host larger amounts of load and DG. This means that the networks are

operated closer to their limits and the safety margins are smaller than before. In order to avoid violating network operational limits (e.g. node voltage) or physical limits (e.g. line current), the automation system needs to monitor the state of the network more closely and take actions if the limits are approached.

In smart grids, the number of real-time measurements is larger than in conventional distribution networks. However, it is still not economically viable to monitor every single network node – including low voltage network nodes – in real-time and this is why distribution system state estimation (DSSE) is needed. In DSSE, the main challenge is to find the most likely state of the network when there is a limited amount of information. At present, the DSSE relies mainly on real-time measurements available from the primary substations and the loads are modelled with load profiles, which are used as artificial measurement (a.k.a. pseudo-measurements). Although the introduction of smart grids and affordable current and voltage sensors will increase the number of real-time measurements, there will still be a need for load profiles, especially in low voltage (LV) network state estimation. DG and active network control are spreading also to the LV side (Repo et al. 2011) and this creates demand for LV network state estimation.

In literature, it has been widely acknowledged that accurate DSSE is needed to enable active network control functions at the core of the smart grid concept. This has been addressed by developing countless new state estimation methods suitable for estimating distribution network states. Weighted least squares (WLS) approach is the most common method utilized in DSSE and it has many variations. Either node voltages or branch currents can be selected as state variables, network can be treated as a whole or divided into measurement areas, or machine learning algorithms can be combined with the WLS method, see for example (Baran & Kelley 1994; Baran & Kelley 1995; Džafić et al. 2013; Wu et al. 2013; Hayes et al. 2015). Also, the possibility to have more real-time measurements has been considered and the best locations for these additional measurements have been analysed, see for example (Baran et al. 1996; Shafiu et al. 2005; Nusrat et al. 2012; Abdel-Majeed et al. 2013; Damavandi et al. 2015; Vasudevan et al. 2015; Xygkis et al. 2016).

The fact that DSSE accuracy can be enhanced by improving the pseudo-measurements has been recognized (Cobelo et al. 2007), but the existing studies have concentrated either on replacing them with real-time smart meter measurements (Baran & McDermott 2009; Abdel-Majeed & Braun 2012; Jia et al. 2013; Alimardani et al. 2015) or using previous day smart meter measurements and short-term forecasting algorithms to supply new pseudo-measurements (Chen et al. 2014; Hayes et al. 2015). The possibility to use smart meter data and classical load research to improve the pseudo-measurements has been largely ignored in literature. This would be a more cost-efficient and practical solution as the real-time reading of smart meters has several challenges; the infrastructure for wide-scale real-time reading does not exist yet, the reading intervals are relatively long, the delays in data transfer are long and unequal, and the reliability of the real-time data is sometimes poor. Also, this would be computationally less expensive than the forecasting based solutions where the forecasts are updated every time new AMR data arrives.

1.2 Smart meter rollout and the ensuing opportunities

In Europe, smart metering is seen as an essential tool for market liberalization, smart grid development and energy saving. The European Parliament and Council directive (2012/27/EU) urges member states to implement remote reading, if the cost-benefit analysis is positive, and at least 14 countries are committed to installing remotely readable electricity meters by 2020. Several countries, including Finland, have already completed the smart meter rollout. It is estimated that in total almost 200 million smart meters will be installed in EU by 2020 (EC JRC 2016). In general, smart meters supply hourly or more frequent interval data on electricity consumption and are remotely readable. The same is true for the previous generation metering system we used to call automatic meter reading (AMR). Both AMR and smart metering systems can supply the interval data utilized in this thesis. Thus in this thesis, they are seen as interchangeable data sources (see Section 2.1 for detailed description of metering systems).

Before AMR and smart meters, the collection of electricity consumption data was very laborious and often the most time consuming part of a load research project. Now, with the above-mentioned meters in place, we can say that "half of the work" has already been done and we are left with the task of analyzing the measurement data.

The literature knows countless studies where AMR data has been used to calculate customer class load profiles. The classification of customers is often made with the help of a clustering algorithm, and many different algorithms have been used successfully, see for example (Chicco et al. 2005; Prahastono et al. 2007; Flath et al. 2012; Haben et al. 2016; Li et al. 2016). The purpose of these studies has often been to produce load profiles for tariff design, market strategy planning and balance settlement purposes. In smart grids, better load profiles are needed also for distribution system state estimation, planning and load forecasting. These less studied applications are the focus in this thesis. Particularly the state estimation and how it can benefit from AMR-based load profiles. Figure 1.1 summarizes the above discussed needs and possibilities, and positions this thesis.

1.3 The evolution and scope of the thesis

The work towards this thesis started in 2007 as a part of the "Methods for Active Distribution Management (ElDig2_VPP, 2006–2008)" project. At that time, active voltage control in distribution networks was studied in the Tampere University of Technology (Kulmala 2014), and it was recognized that an accurate distribution system state estimator is needed to complement the developed voltage control method. The author's M.Sc. thesis (Mutanen 2008) studied the usage of remotely readable measurements in distribution system state estimation. During this research, the author observed that the DSSE accuracy could be improved by using new DSSE methods, by adding real-time measurements, by optimizing the real-time measurement locations, and by improving the pseudo-measurements with the help of AMR measurements. The latter approach was selected for further development because, at the time, AMR was a hot discussion topic in Finland. The first large-scale AMR implementations were under way and a majority of the distribution system operators (DSOs) had decided to invest in AMR (Kirjavainen & Seppälä 2007). The AMR-based load profiling was studied in the

"Interactive Customer Gateway for Electricity Distribution Management, Electricity Market, and Services for Energy Efficiency (INCA, 2008–2010)" and "Smart Grids and Energy Markets (SGEM, 2010–2016)" projects, while at the same time DSSE development was continued in the "Active Distribution Network (ADINE, 2007–2010)", "Intelligent Electrical Grid Sensor Communications (INTEGRIS, 2010–2013)", and "Ideal Grid for All (IDE4L, 2013–2016)" projects.

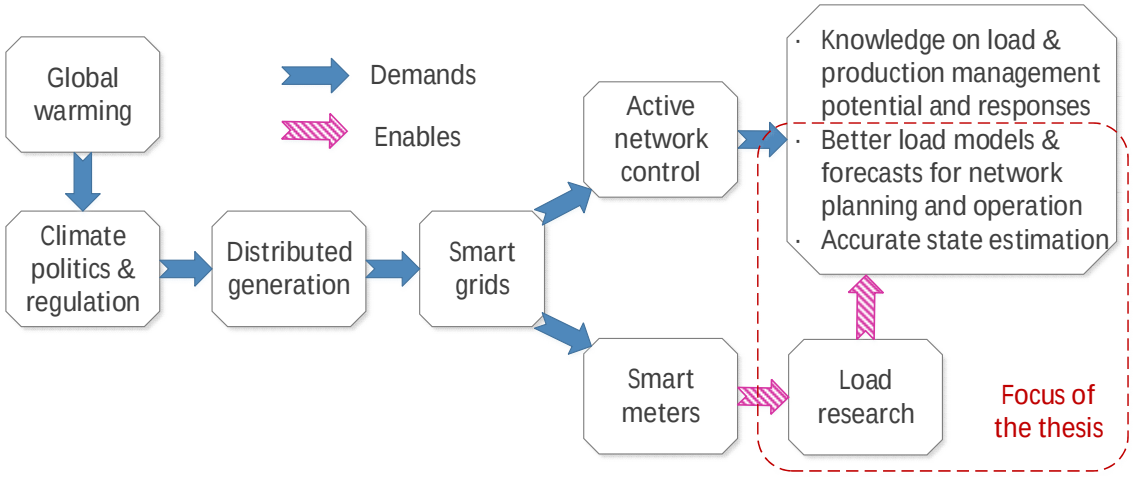


Figure 1.1 The chain of demands leading to the field of this thesis.

The domestic INCA and SGEM projects were carried out in close co-operation with electricity retailers, DSOs and industry operating in the field of electricity distribution. In order to ensure fast and straightforward application of the developed load profiles, it was decided that the basic structure of the existing Finish load profiles (which is described in Subsection 2.2.1) would be kept unchanged and only the content of the load profiles would be updated with the help of AMR measurements. This principle was followed throughout this thesis, except in publication [P7], which was written during the authors' research exchange visit and takes a British point of view to load profiling. The focus of the EU funded ADINE, INTEGRIS, and IDE4L projects was on demonstrations and the DSSE development done in these projects concentrated on fulfilling the needs of active network control algorithms. Also, since the demonstrations were done in real distribution networks, DSSE robustness and ability to estimate different types of networks with different measurement configurations was emphasized.

As shown in Figure 1.1, this thesis focuses on load research and application of load research results in distribution network analysis. The research questions this thesis aims to answer can be summarized as follows:

- How AMR measurements can be used to improve the load profiles?
- How the customer classification can be improved and automated?
- How the new AMR-based load profiles improve distribution network analysis, especially medium and low voltage network state estimation?

There are many different ways in which AMR measurements can be used to improve the existing load profiling practices. These methods are discussed in Chapter 3. One of the most interesting methods is the use of clustering algorithms in customer classification,

which due to its importance and complexity has been separated as its own task and is discussed in Chapter 4. There are countless applications for the new AMR-based load profiles, as shown in Figure 1.2, but only a few of those are studied in this thesis. In [P6], the effect of new load profiles on distribution network peak load modelling is shown. In [P8], the new load profiles are used for load forecasting. In [P9], the new load profiles are used to improve the DSSE accuracy. The introductory part of this thesis emphasizes the latter application as improved distribution network load flow calculation and DSSE were the primary motives behind the load profiling efforts. Chapter 5 reviews the developed DSSE method and discusses the effect the new AMR-based load profiles have on the DSSE accuracy.

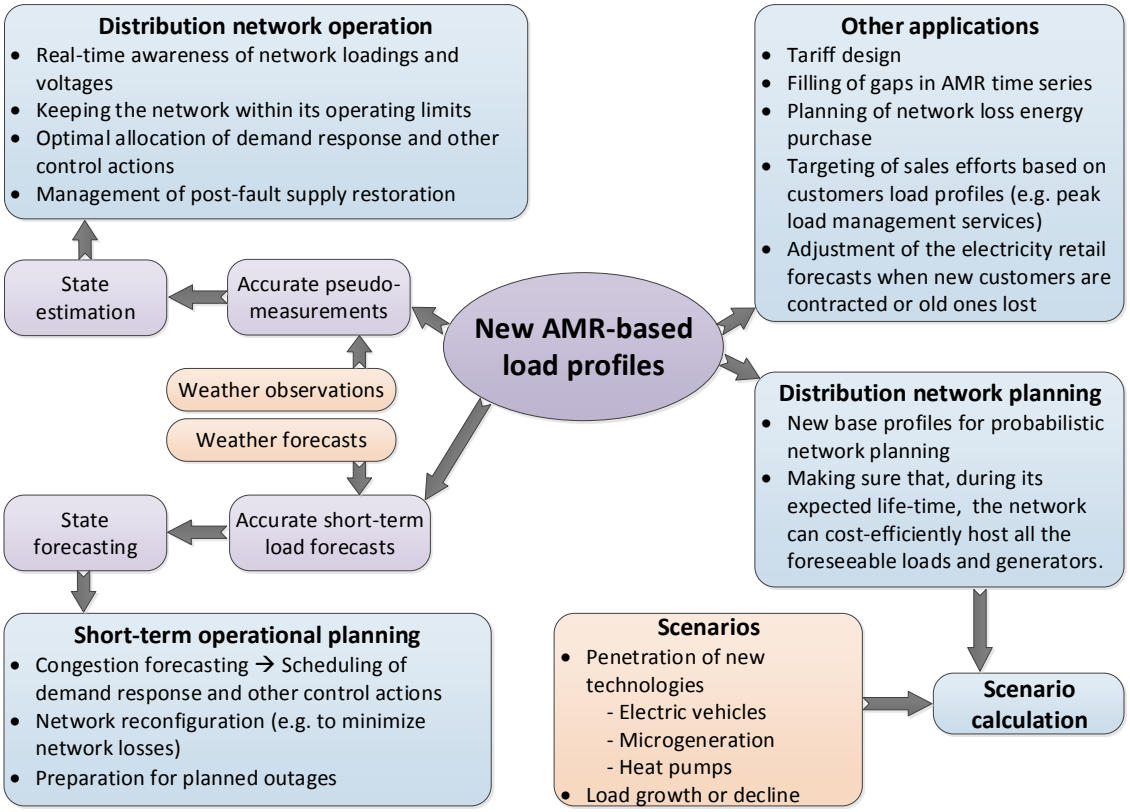


Figure 1.2 Applications for the new AMR-based load profiles.

In addition to work presented in this thesis, the author has also contributed to studies supporting the methods developed here. Demand response (DR), changes in customer behaviour, and technological development will change how electric loads behave and how they should be modelled. The author has supervised a M.Sc. thesis studying the effects of DR and microgeneration on load profiling (Grip 2013) and contributed to the following publication (Grip et al. 2014). The author has also supervised a M.Sc. thesis studying customer behaviour change detection based on AMR measurements (Chen 2014) and contributed to the following publication (Chen et al. 2015). The research on DR modelling and AMR-based change detection are continued in the "Improved Modelling of Electric Loads for Enabling Demand Response by Applying Physical and Data-Driven Models (RESPONSE, 2015–2018)" project. The detailed results of these

studies have been excluded from this thesis in order to maintain a coherent and well-outlined structure.

The load research material used in this thesis consists mainly of hourly interval data measured from Finnish end users. In some other countries, half- or quarter-hourly interval data may be available but this does not change the principles or prevent the application of the developed load profiling methods. However, the differences in load profile formats must be taken into account. The Finnish load profiles cover the whole year, while in most other countries typical daily profiles (TDPs) are preferred. The usage of TDPs is studied in publication [P7]. In this thesis, Matlab is the primary tool used in load research and DSSE development and analysis. Matlab can easily handle AMR data sets containing thousands of customers and provides efficient matrix operations needed in DSSE calculation. The upper limit of the AMR data set size ranges from tens to hundreds of thousands customers depending on the size of the computer main memory, time series length, and interval length. The developed DSSE method was tested with Matlab and Real-Time Digital Simulator (RTDS) simulations, and with demonstrations done in real distribution networks. Especially the IDE4L project contained many real-life demonstrations, but the results of those are left to lesser attention in this thesis. In real-life demonstrations, the number of available reference measurements is often low. Moreover, uncertainties in input parameters and measurements make identification of the error sources difficult in real-life demonstrations.

1.4 Main contributions

The main contributions of this thesis are:

- The benefits of using AMR data in load profiling were shown by comparing new AMR-based load profiles with the existing customer class load profiles. AMR measurements were used for updating the existing customer class load profiles, customer reclassification, clustering, and individual load profiling.
- A two-stage clustering method for clustering electricity customers was developed. The method starts from the raw AMR data and outputs cluster profiles, new customer classification, and individual load profiles for large and abnormally behaving customers.
- The applicability of 15 clustering algorithms for electricity customer clustering was tested. The best algorithms were compared and sensitivity analyses were performed.
- A DSSE algorithm for smart grid environment was developed and tested in real-life demonstrations. The developed DSSE algorithm is able to use all types of conventional real-time measurements (phasor measurement units are excluded), calculate weakly meshed distribution networks, provide uncertainties for the estimated states, and operate in a decentralized manner.
- It was proven, through simulations with real AMR data and a real distribution network model, that the developed AMR-based load profiles improve the DSSE accuracy. The simulations were performed with the DSSE algorithm developed in

this thesis but the simulation set-up was such that the results are generalizable to most WLS estimators.

1.5 Publications and author's contribution

This thesis includes nine publications that represent original work in which the thesis author has been an essential contributor. Publications [P1]–[P3] discuss distribution system state estimation and publications [P4]–[P8] discuss AMR-based load profiling and forecasting. Finally, publication [P9] combines AMR-based load profiling and distribution system state estimation. Apart from publications [P7] and [P8], the thesis author has been the corresponding author and has been solely responsible for writing and editing the publications. Prof. Pertti Järventausta and Prof. Sami Repo have been the supervisors of this dissertation work and have contributed to the publications through guidance during the research work and by commenting on the publications prior to publishing. The roles and contributions of the other co-authors have been described in the list below.

- Publication [P1] is based on the author's M.Sc. thesis and discusses how all typically available real-time measurements could be utilized in distribution system state estimation. A WLS-based DSSE algorithm is proposed and simulation results using the IEEE 37-bus test feeder are presented. All the work presented in this publication has been done by the author.
- In publication [P2], the author added dad data detection to the DSSE algorithm presented in [P1] and tested the revised algorithm in Matlab, real-time digital simulator (RTDS), and real distribution network. M.Sc. Antti Koto and Ph.D. (tech) Anna Kulmala participated in RTDS simulations and real-life demonstration. Antti Koto and was responsible for implementing the data transfer between Matlab, supervisory control and data acquisition (SCADA), and RSCAD. Anna Kulmala was responsible for implementing the coordinated voltage control (CVC) algorithm tested in conjunction with the DSSE algorithm, the results of which are presented in separate publications (Kulmala et al. 2010) and (Kulmala et al. 2012). Analysis of the results and writing was done solely by the author.
- In publication [P3], the effect of input measurement reading frequency and averaging time on DSSE accuracy is tested with RTDS simulations. The simulation results in this publication are based on Atte Löf's M.Sc. thesis where he tested the DSSE algorithm developed by the author. Prof. Sami Repo and the author defined the used accuracy metrics and outlined the simulation plan. The author supplemented M.Sc. Atte Löf's analyses on the simulation results, added the parts relating to the real-life demonstration, and wrote the publication. Ph.D. Davide Della Giustina commented on the publication prior to publishing.
- Publication [P4] presents and compares methods for utilizing AMR measurements in customer classification and load profiling. Customer reclassification, load profile updating, clustering and individual load profiling are studied and compared. All the work presented in this publication has been done by the author.

- Publication [P5] proposes a customer classification and load profiling method that includes load temperature dependency modelling, outlier filtering and a clustering algorithm that is based on iterative self-organizing data-analysis techniques (ISODATA). The method presented in this paper is in great measure based on the previously unpublished work of M.Sc. Maija Ruska. With permission and some help from Maija Ruska, the author recreated, tested and published the method she had initially developed when working at the VTT Technical Research Centre of Finland.
- Publication [P6] shows how AMR-based load profiles improve traditional network analysis that utilizes confidence levels. The author developed a method for updating existing load profiles and for creating cluster profiles. The author then compared these new profiles with standard customer class load profiles. M.Sc.'s Matti Kärenlampi and Pentti Juuti from ABB (ASEA Brown Boveri) supplied the prototype version of MicroSCADA Pro DMS 600 distribution management system that the author used when demonstrating how cluster profiles can be used side by side with old and updated customer class load profiles.
- In publication [P7], Gaussian mixtures and mixtures of factor analyzers are used to cluster and model residential customers. The identified load models are compared to standard load profiles and their benefits are demonstrated using statistical load flow. The author wrote this paper together with Ph.D. Bruce Stephen with fifty-fifty contribution. Bruce Stephen was the corresponding author and wrote the introduction, conclusions, and theoretical parts containing equations, while the author wrote the results chapter and parts describing the load profiling practices and status of the metering systems. The rest of the publication was written with mixed contributions. The author did also much of the practical work; coding, calculation of the results, and figure drawing. Ph.D. Stuart Galloway and Prof. Graeme Burt commented on the publication prior to publishing.
- Publication [P8] assesses how AMR-based load profiles, neural networks (NN), and Kalman-filter based predictors with input nonlinearities are suited for short-term load forecasting. This paper was written together with Ph.D. (tech) Pekka Koponen and Ph.D. Harri Niska with approximately equal contributions. Pekka Koponen was the corresponding author and made the Kalman filter based predictor. The author supplied the input data, made the AMR-based load profiles and wrote the accuracy calculation script. Harri Niska made the NN model. Each author contributed to writing by describing the forecasting method they had developed. Pekka Koponen compiled the texts and wrote the first draft, which the others then helped to finalize.
- In publication [P9], different AMR-based load profiling methods are compared and their effect on the DSSE accuracy is evaluated. The accuracy evaluation is done through a case study where a real distribution system is simulated as accurately as possible using real network data and real measured loads. All the work presented in this publication has been done by the author.

2 Background to load profiling

This chapter provides background information necessary for understanding the environment in which this thesis has been written. The purpose of this chapter is not to provide a comprehensive state-of-the-art literature review. Instead, literature reviews on individual research topics are presented in later chapters.

2.1 Electricity meter reading

Meter reading is an essential part of electricity distribution network and electricity retail businesses. Earlier, when only analog electricity meters were used, meter reading was very labor-intensive and was therefore done infrequently. Customers were regularly billed based on their estimated electricity consumption and balancing bills were sent when the meters were finally read. Also, the detection of low voltage network faults was slow as it relied heavily on customer complaints received via telephone.

Nowadays, automatic meter reading systems collect consumption, diagnostic and status data from electronic energy meters and transfer this data automatically to a central database for billing, troubleshooting and analyzing. This new digital technology eliminates the need for on-site meter reading, reduces unnecessary visits to the metering site, and accelerates both electricity distribution and retail businesses.

Meter reading systems have evolved over the years, and so have their names. As new functionalities have been added to the meter reading systems, the naming used in product brochures and scientific publications has changes from automatic meter reading to smart metering. In the introductory part of this thesis, the earlier term AMR is used. Although limited to one-way communication, the AMR system is able to provide all the measurement information used in this thesis. The attached publications also feature other terms such as smart metering and smart meters. Next, short descriptions for different metering systems are given.

2.1.1 Automatic meter reading

Automatic meter reading (AMR) system collects data from metering points (electricity, water or gas) via one-way communication. In some early implementations, this meant collection of monthly energies through short-range communication devises that required either an on-site visit or a drive-by. Nowadays it is common that the AMR system automatically transfers the data, which can contain also interval data on consumption, to a central database as often as once a day using either wireless (radio frequency, mobile phone network, wireless local area network etc.), wired (power-line communication, fiber optics, telephone cables etc.), or combination of wireless and wired communication technologies. The main advantage of AMR is that the on-site meter reading is not needed and billing can be based on actual rather than estimated consumption. The interval data collected with AMR can also be used in load research as has been done in this thesis. Compared with other more advanced metering techniques, the most defining character of the AMR systems is their mainly unidirectional flow of information. However, this does

not prevent AMR meters from sending locally initiated alarms such as power outage or bad power quality notifications.

2.1.2 Advanced metering

Advanced metering generally refers to the next generation metering solutions that allow bidirectional flow of information. When talking about advanced metering, terms such as advanced metering management (AMM), advanced metering infrastructure (AMI), smart metering infrastructure (SMI), and smart metering are often used interchangeably although one could argue that they have some differences. Some see that AMM includes all the traditional AMR features and adds new functionalities that utilize two-way communication to control the metering system and the distribution network, but excludes the hardware and software that is needed to implement the two-way communication. The AMM enabling infrastructure is thus considered separate and is termed either as AMI (PE EPS 2012) or SMI. Then again, there are also those who think that AMM is part of AMI (or SMI), which is used to describe the whole advanced metering system (Vayá et al. 2016). Smart metering is an even more obscure term, but in general it appears to contain the same properties as AMM and is often used as a synonym for advanced metering and AMI (Koponen et al. 2008).

In this thesis, AMI, SMI and smart metering are bundled together into a system that is assumed to also contain AMM. Figure 2.1 shows how these metering terms overlap. The boundaries between the terms are often fuzzy and there are some exceptions. For example in Finland, the MELKO system enabled bidirectional data transmission and load management, in addition to remote meter reading, already in the mid-1980s (Kosonen 2008), long before the introduction of advanced metering. Moreover, the latest generation AMR solutions already included many of the functionalities nowadays associated with AMM, AMI, SMI and smart metering. Although the advanced metering systems include all the AMR functionalities—or better versions of them—the AMR systems do not include all the advanced metering functionalities.

In some cases, the advanced metering systems can be configured so that the meters are read several times per hour (e.g. every 5–15 minutes). In this thesis, measurements with this kind of reading frequency are considered to be real-time measurements.

2.2 History of load profiling

Knowing the load magnitude and its temporal variation has always been vital for electric power industry. This need is not limited only to the total load, but also the sub-loads from which the total load is composed are important. In distribution network operation and planning, and electricity retail, load profiles describing the behavior of typical customers in different customer classes (e.g. industry, commerce, and housing) are needed. The forming of such profiles is called *load profiling*, which is an important sub-field in *load research*.

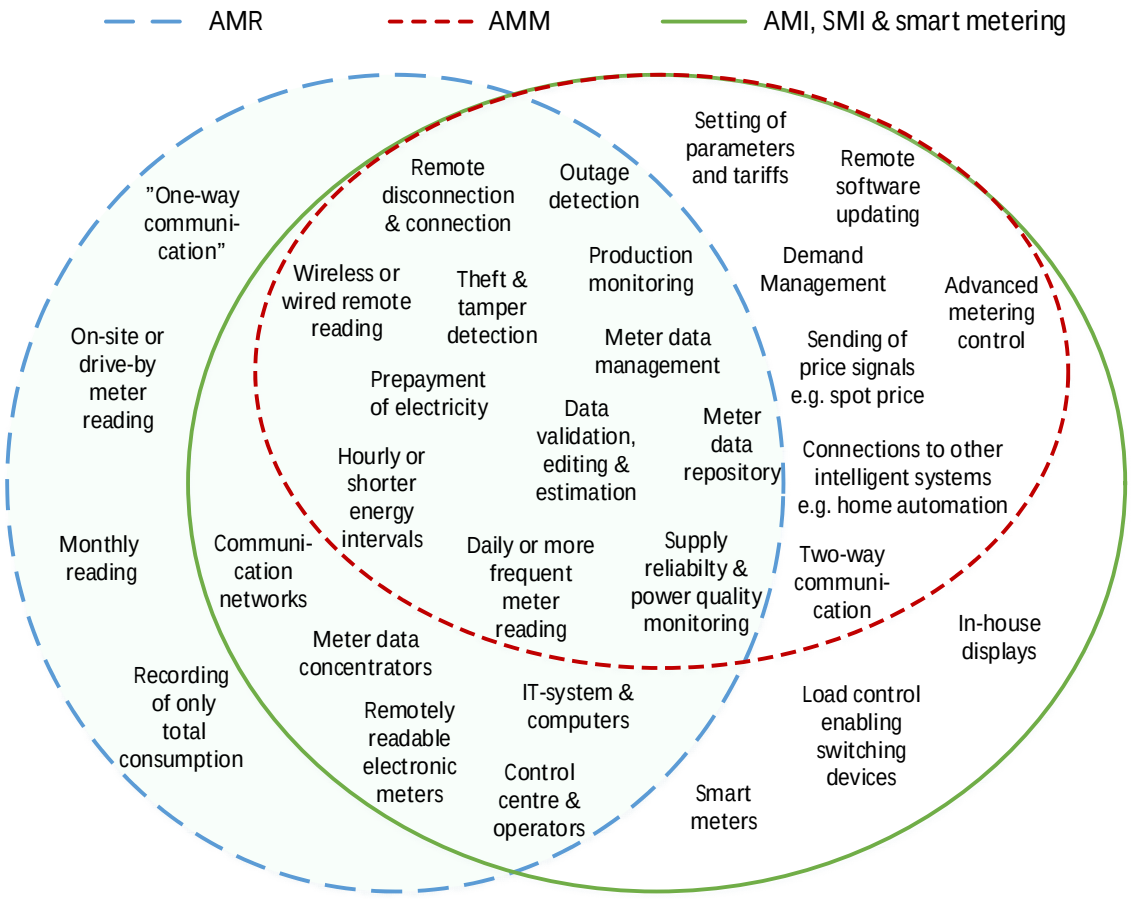


Figure 2.1 Metering scene and the authors view on metering system categorization.

The demand for load profiles shot up when the electricity supply markets were opened for competition. In open electricity markets, retailers need to optimize their product portfolios and minimize the risks stemming from the fluctuating electricity prices. In the 1990s, United Kingdom and the Nordic countries were in the forefront of the electricity market liberalization (Kopsakangas-Savolainen 2002). At that time, very few customers had remotely readable interval meters and customer class load profiles were needed for tariff design, balance settlement, and targeting of sales efforts.

Customer class load profiles are also needed in load forecasting, distribution network load flow analysis, and state estimation. These form a basis for network operation, dimensioning and design. In recent years, the importance of these applications has increased as smart grids have emerged. In smart grids, the above-mentioned tasks must be done carefully and considering the expected loads, physical network limits and capabilities of the active resources.

Before AMR, load profiling was done primarily by measuring a sample of end users, categorizing them by the type of electricity use, and generalizing the achieved results to cover other customers of the same type. In academic literature, other methods such as the bottom-up approach are occasionally used (Bizzozero et al. 2016). The bottom-up approach starts from the electric footprint of individual appliances and combines this with information on appliance penetration statistics, typical appliance usage, the number of

occupants (in residential buildings), and other available information. Actual consumption measurements can also be used in validation and tuning of the bottom-up models.

Next, the history and present state of load profiling in Finland and some other countries is reviewed. In all the studied countries, the load profiles are based on measurement samples.

2.2.1 History and present state of load profiling in Finland

One of the earliest documented load research projects in Finland dates back to the 1950s when the Helsinki municipal electric utility studied how the total electric load is divided into sub-loads each with a distinct load profile (Puromäki 1959a & 1959b). Although the basics of load research had already been well understood, the studies were restricted by the technology of the time. In the absence of load recording devices, the hourly average loads had to be logged manually and this naturally limited the amount of data collected. Also, the calculation of load profiles was very laborious using only tabulating machines and mechanical desktop calculators (Puromäki 1959a).

In 1983, Finnish electric utilities started a large-scale co-operation in load research. Over 40 utilities joined the load research project coordinated by the Association of Finnish Electric Utilities (in Finnish: Sähkölaitosyhdistys, abbr. SLY, changed its name later to *Sener* and merged with Finnish Energy). During this project, hourly electricity consumption measurements from over 1000 customers were collected (Sener 1992). After the first measurement period (1983–1985), customer class load profiles for 18 customer classes where published (SLY 1986). The second measurement period (1986–1988) concentrated on industry and service class customers and after the results had been analyzed, Sener was finally able to publish load profiles for 46 different customer classes (Sener 1992). These include customer class load profiles for housing, agriculture, industry, commerce and administration each divided into several sub-classes according to their electricity use pattern defining characters (building type, heating solution, field of business, number of work shifts etc.). Since publication, these load profiles have been widely used in Finnish distribution network load flow calculation, state estimation, network planning and tariff planning (Seppälä 1996).

The customer class load profiles that resulted from the 1983–1994 load research study (from now on referred to as Sener profiles), are still the only comprehensive set of load profiles publicly available in Finland. After 1994, the responsibility for load research was given to VTT Technical Research Centre of Finland. VTT Technical Research Centre of Finland developed the load profiling methodology further, updated some load profiles and defined load profiles for two new customer classes; green houses and three-shift industry (Jalonen et al. 2003). However, these new load profiles are available only to those 15 companies that participated in this project. After this, load research efforts have been limited to company scale studies where they have defined new customer class load profiles for their own use. Some companies have also created individual load profiles for their largest customers.

Despite some deficiencies (q.v. Subsection 2.3.2), the structure and usage of the load profiles have not changed since the Sener profiles were introduced about 25 years ago. In Finnish DSOs, each individual customer is classified into one of the existing customer classes by using the information available in the customer information system (CIS). CISs contain information on each customer's network connection, type and electricity consumption. Optionally, some of the largest customers can be modelled with their own individual profiles. All the customers are linked to the geographic network model in the network information system (NIS). This enables network calculations using the load profiles.

The load profile structure used by most Finnish DSOs' software applications represents the expected value and standard deviation for the customer's hourly load as a linear function of the annual energy consumption. The load profiles can be represented either as topography or as index series. In topography, the expected value and standard deviation for hourly load are given for every hour of the year. Expected value and standard deviation are usually given for a base energy consumption of 10 MWh/year and when applied, the values are scaled so that the sum of the expected values correspond with customer's annual energy. The index series model the yearly energy consumption pattern in a more compact form. The index series are composed of two parts; yearly indices and daily indices. The yearly indices model seasonal variation with 26 two-week indices and the daily indices model hourly variation during three different day types (working day, Saturday and Sunday) separately for each two-week period. The index series contain expected values for both yearly and daily indices but the standard deviations, which are given as a percentage of the expected value, are defined only for the daily indices. One index series thus contains $26+3\times24\times26=1898$ parameters for load expected value and $3\times24\times26=1872$ parameters for load standard deviation. Topographies consider special holidays, but in the index series public holidays and eves are modelled as Sundays and Saturdays respectively. Both in topographies and in index series the hourly reactive powers are calculated using one customer class specific power factor. (Sener 1992; SLY 1992)

After market liberalization, load profiles were used also in balance settlement. Nowadays, the balance settlement is done mainly with AMR measurements (Finnish Energy 2016) and load profiles are used only for those few customers that are not within interval metering. The load profiles used in the balance settlement are different from the ones used in the network calculation. In the balance settlement, customers are divided into three customer classes and only three load profiles are used. The customer classes are: households with electricity consumption equal or less than 10 MWh/year, households with electricity consumption greater than 10 MWh/year, and others unmeasured customers. (REG 1.3.2009/66)

2.2.2 History and present state of load profiling in some other countries

Sweden has a long history in load research, dating all the way back to the 1940s when Sten Velander formulated the relationship between peak power and annual energy

consumption (Neimane 2001). Velander's formula has since been widely used in distribution network dimensioning. Outside Scandinavia, this method for transforming annual energies into peak power is sometimes known as the Strand-Axelsson formula (Provoost 2011). When applied to a large homogeneous group of electricity users with correct parameters, Velander's formula can be quite accurate. However, when calculating peak power for a group that contains customers from several customer groups, which do not peak at the same time, Velander's formula can result in too high values (Neimane 2001).

To address the problems associated with Velander's formula, Swedish Association of Electric Utilities (Svenska Elverksföreningen, nowadays part of Elforsk) conducted a study where they measured electricity consumption from 400 electricity customers with 15-minute intervals for a period of one year. After analysis, typical daily profiles (TDPs) for roughly 40 customer groups were published in 1991. This set of profiles covers domestic, commercial and industrial customers. The format of load profiles is such that 16 TDPs exist for each customer group. Separate profiles have been defined for working days and non-working days in three different seasons (winter, spring/autumn, and summer) and in different outdoor temperatures (three for winter and spring/autumn, and two for summer). Load standard deviation is also presented in the load profiles. (Engblom & Ueda 2008; Dahlström et. al. 2011; Hemmingson & Lexholm 2013)

After 1991, several efforts have been made to increase the knowledge on electricity consumption temporal behavior and temperature dependency. Corfitz Norén and Jurek Pyrko have studied electricity consumption in schools, hotels, grocery stores and nursing homes (Norén 1997; Norén & Pyrko 1998a; Norén & Pyrko 1998b; Norén & Pyrko 1999). Elforsk has published studies on electricity consumption in very cold temperatures (Larsson et. al. 2006; Dahlström et. al. 2011). In these studies, TDPs for different types of residential customers were calculated in different outdoor temperatures.

Despite the above-mentioned efforts to improve the load profiling accuracy, it is possible that the original load profiles from 1991 are still used in some electricity companies and commercial software. For example, Mälarenergi Elnät AB uses the load profile package *Betty 1.2*, which seems to coincide with the 1991 load profiles, in their NIS (Arvidsson 2015). The Betty load profile package is also used in MarkedMath Europe AB's Pluto pricing tool (MarketMath 2016).

In **United Kingdom (U.K.)**, coordinated load research has been practiced since the 1950s when the first Electricity Council load research program started. At the beginning, load profiles were needed mainly for designing and setting retail tariffs (Allera et. al. 1990). However, when the electricity supply markets were liberalized, a new need for load profiles arose. In liberalized energy markets, a distribution network can contain customers supplied by different electricity suppliers. In this case, there must be a way to quantify how much energy the customers of each supplier have used during each half hour interval (in U.K.). The amount of energy the electricity suppliers purchase and the amount of energy their customers consume should match in each half-hour period. The electricity settlement process enforces this and charges the suppliers for any imbalance. It was

decided that to avoid installation of new half-hourly meters, customers below 100 kW maximum demand would be settled using load profiles. (ELEXON 2013)

Load profiles for eight different customer classes were defined. These customer classes cover domestic and non-domestic customers with and without time-of-use tariff and non-domestic maximum demand customers with four different peak load factors. Typical daily profiles (TDPs) containing 48 half-hourly usage levels for three different day types (working day, Saturday, and Sunday) in five different seasons (winter, spring, summer, high summer, and autumn) were defined for each customer class. The TDPs are calculated from measurement samples collected all around the U.K. using multiple linear regression that takes into account weighted outdoor temperature from three previous days, sunset time, and day of the week. In their standard form, the TDPs are given in long-term average temperature but when the company currently responsible for management and development of TDPs (ELEXON Ltd.) sends the daily-calculated profiles to balance responsible suppliers, all the above-mentioned regressors are taken into account. (ELEXON 2013)

In **Germany**, standard load profiles (SLPs) were introduced in the 1980s and have since been used in tariff planning, grid planning, and consumption forecasting. For example, the energy consumption forecasts of a DSO balancing group are usually made with SLPs. Standard load profiles for nine customer classes (1×domestic, 6×industrial and 2×agricultural) have been defined by the Federal Association of Energy and Water Management (Bundesverband der Energie- und Wasserwirtschaft). Each SLP describes typical daily loading in three different day types (workday, Saturday and Sunday) and in three time intervals (winter, summer and a transition period containing both spring and autumn). Each daily profile contains 96 values. (Abdel-Majeed 2016)

Customers with energy consumption larger than 100 MWh/year are metered remotely and the metered quarter-hourly consumption data is used in balance settlement. Smaller customers are settled with SLPs. In addition to the above presented nine basic SLPs, the German balance settlement practices allow the usage of DSO specific load profiles and many utilities have created additional profiles for telecommunication towers, street lighting, photovoltaic plants, storage heaters, heat pumps etc. Some of these additional profiles are temperature dependent and have been defined in different temperatures.

2.3 Load profiles in distribution network calculation

Load profiles are widely used in electric power industry. They are needed in network operation and planning, tariff design and production planning. They are essential in applications that require knowledge on the end user loads, for example in load flow calculation, which is one of the basic functions in distribution network analysis. Load flow calculation is necessary for determining the line current flows, node voltages and power losses. Load profiles are also needed in other applications such as distribution system state estimation and distribution transformer load management. The value and usability of the load profile derived calculation results increase if they are combined with geographical network information and are drawn over a background map. For example,

the network component loadings can be dawn with different colours so that the operator can see the overall loading situation at a glance.

2.3.1 Distribution network calculation in Finland

In Finland, distribution companies have long experience in using geographical network information systems and load profile based network calculation. The first network information systems were brought into use already in the 1980s and they included network documentation, map drawing, and applications for network planning and calculation (Järventausta et al. 2011). Both medium voltage (MV) and low voltage networks are modelled within NIS and all customers, even individual LV customers, are connected to the network model. Parallel with the Finnish load research project, applications for network load computation, network planning, and electricity pricing were developed and by the 1990s several Finnish software companies had produced commercial NIS and load flow calculation software products that utilize load profiles (Seppälä 1996).

Nowadays, load flow calculation with load profiles is routine for DSOs. In the NIS, the calculation starts from individual customers and propagates upwards. First, yearly energy estimates are fetched from the CIS and the customer class load profiles are scaled to match each customer's yearly energy. After this, estimates for load expected and standard deviation values for every hour of the year are known. When higher level loadings (for example trunk line or transformer power flows) are needed, the customer level loads are aggregated according to the probability theory. For simplicity, loads are assumed normally distributed and independent. In that case, the aggregated load expected values $\mathrm{E}[P_{ag}(t)]$ and standard deviations $s_{ag}(t)$ for n customers should be calculated with:

$$\mathrm{E}\left[P_{ag}(t)\right] = \mathrm{E}[P_1(t)] + \mathrm{E}[P_2(t)] + \cdots + \mathrm{E}[P_n(t)] \quad (1)$$

$$s_{ag}(t) = \sqrt{s_1(t)^2 + s_2(t)^2 + \cdots + s_n(t)^2}, \quad (2)$$

where $\mathrm{E}[P_i(t)]$ is the expected value of customer i active power during time t and $s_i(t)$ is the standard deviation of customer i (active power) during time t (Seppälä 1996). In practice, individual loads are not normally distributed. However, since sums of many individual loads are often needed in distribution network calculation, and since the central limit theorem states that the distribution of the sum of many independent random variables tends toward a normal distribution even if the underlying variables are not normally distributed, the assumption of load normality is reasonable. In addition, the assumption on the load independence can be strengthened by modelling the correlation-causing factors, such as the load temperature dependency, separately.

The stochastic nature of the loads is taken into account when calculating peak loads. Load values with different excess probability levels are used in distribution network calculation. The load $P_p(t)$ having an excess probability of p % can be calculated with:

$$P_p(t) = \mathrm{E}[P(t)] + z_p \times s(t), \quad (3)$$

where z_p is the Z-score corresponding to excess probability p. The load values with excess probability of around 10 % are relevant for voltage drop calculation, while smaller probabilities are used when studying loading limits. The load expected values (50 % excess probability) are used when calculating network losses. (Lakervi & Holmes 2003)

The statistical properties of the load profiles can be utilized in probabilistic load flow calculation where the line current flows and voltage drops are determined with a certain excess probability. The probabilistic load flow can be based, for example, on the backward/forward sweep method illustrated in Figure 2.2. The steps in this figure are:

1. Calculate $P_p(t)$ for nodes C and D using (3).
2. Calculate currents $\bar{I}_{BC}(t)$ and $\bar{I}_{BD}(t)$ with an equation $\bar{I} = \bar{S}^*/\bar{V}^*$ derived from the basic power equation. Then, calculate the power losses $\bar{S}_{h\_BC}(t)$ and $\bar{S}_{h\_BD}(t)$ with an equation $\bar{S}_h = |\bar{I}|^2 \times \bar{Z}$ (reactive loads and reactive losses are ignored in this simple example, and therefore $\bar{S} = P$ and $\bar{S}_h = P_h$)
3. Calculate $P_p(t)$ for aggregated power in node B using (1), (2), and (3). Then, add to this aggregated power the power losses calculated in step 2.
4. Calculate current $\bar{I}_{AB}(t)$ as in step 2 (note that the Kirchhoff's current law does not apply in probabilistic load flow calculation).
5. Assuming the node A voltage is known, calculate the node B voltage with an equation $\bar{V}_B = \bar{V}_A - \bar{I}_{AB} \times \bar{Z}_{AB}$.
6. Calculate node voltages $\bar{V}_C$ and $\bar{V}_D$ similarly as $\bar{V}_B$ in step 5.

In steps 2 and 4, the node voltages are not known and need to be replaced with network nominal voltages, which can later be replaced with the voltages calculated in steps 5 and 6. The backward/forward sweep procedure is thus iterative and must be repeated until convergence is achieved. In real applications, reactive powers and line charging currents are naturally also taken into consideration. In practical MV network calculation, the LV network loads are often aggregated directly to the distribution transformer level and the LV network losses are approximated with a constant loss factor.

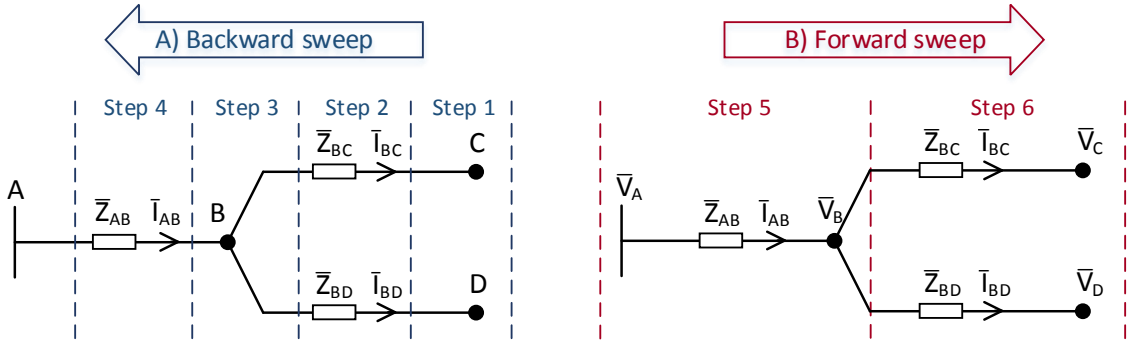


Figure 2.2 Backward/forward sweep method used in probabilistic load flow calculation.

SCADA system provides real-time measurement and switching state information from the distribution network. This information is often limited to measurements and switches located in primary substations and load profiles are needed in DSSE to make the system observable. The real-time SCADA measurements tell how large the substation total load

and feeder loads are, but the load distribution at lower network levels is estimated based on load profiles.

The DSOs have a practice of performing network wide monitoring calculations with load profiles. The purpose of these calculations is to make sure that the network can handle the present or simulated peak loads and find the network sections and components that need reinforcement. In addition to voltage drops and component loadings, also energy losses, interruption costs and short circuit currents are computed and used in investment planning (Lakervi & Holmes 2003). Monitoring that the past states of the network have been acceptable is one of the few applications where previous year AMR measurements could be used directly to replace the load profiles. Even then, the limitations of the directly used AMR data should be considered; the previous year may have been exceptionally warm or cold and subsequently the loads may have been lower or higher than in a normal year.

2.3.2 Defects in the existing load profiles

The above introduced and presently used load profiles have many defects, some more than others. Most of the shortcomings in the existing load profiles can be traced back to the pre-AMR era when load research was expensive and sampling was necessary in load profiling. One of the most prominent issues is the old age of the existing load profiles. For example, in Finland the Sener load profiles are based on measurements done 27–33 years ago and electricity consumption habits have changed considerably over the last decades; heat pumps have become popular, lighting and refrigeration devices have become more energy efficient, the amount of computers and home electronics has skyrocketed, and car indoor heaters have become common (Sener 1992; Adato 2013). Consequently, the actual load profiles have drifted away from the Sener profiles.

In sampling based load profiling, the accuracy of the profiles depends much on the sample sizes. In many earlier load research studies the sample sizes have been insufficient. For example, in Finland the Sener profiles were calculated based on 639 measured time series (many of the original 1000+ measurements were omitted from the final analysis) and the sample sizes varied between two and 65. In Sweden, the 1991 load research study used measurements from 400 electricity customers and divided them into 40 customer classes; this means that on average the sample size was only ten. The literature gives varying numbers for a sufficient sample size. Lakervi and Holmes (2003) recommend at least 100 customers per customer class with consumption records taken over the last three years. Argonne National Laboratory (1980) derives the following formula for the minimum sample size:

$$n = \left(\frac{sz}{r\bar{X}}\right)^2, \tag{4}$$

where s is the sample estimate of the population standard deviation, z is the Z-score determined by the chosen confidence level, r is the chosen reliability level, and $\bar{X}$ is the sample mean. If we assume that standard deviation is 50 % of the sample mean and target 90 % confidence with ±10 % reliability, (4) gives a minimum sample size of 68. Small

samples are also sensitive to classification errors. In a small sample, even one wrongly classified customer can be a source of significant sampling error.

Another substantial error source is the classification of the electricity end-users. The type of customer is usually determined through a questionnaire when the network connection is contracted and is rarely updated afterwards. In reality, the customer type may change, for instance, because of a change in the heating solution, an addition of new type of electric load (such as an electric vehicle), or the change of customer activity. For example, in Finland the number of heat pumps has multiplied during the last decade and the number of farms is decreasing steadily (SULPU 2015; Luke 2016). Figure 2.3 shows how the number of installed heat pumps has grown in Finland. The majority of the installations are air-to-air heat pumps, which are typically used to supplement direct electric heating. This means that there are now many houses that are classified as direct electric heating customers but are actually using a hybrid of direct electric and heat pump heating. Figure 2.4 shows how the number of farms has decreased during the years 1995–2015. This means that there are now many farmhouses that are classified as farms even though the farming activities have ended.

The lack of or defects in the outdoor temperature dependency parameters are also major error sources in many load profiles. In Finland, temperature dependency parameters for Sener load profiles have been published only for January (Sener 1992) and even these are rarely used. Instead, the present industry standard is to use −4 %/°C temperature dependency for customers with electric heating and assume that other customers do not have any temperature dependency. This practise is used for example in Seppälä (2007). The above-described approach is of course very coarse and neglects the fact that also many other types of customers exhibit some degree of temperature dependency.

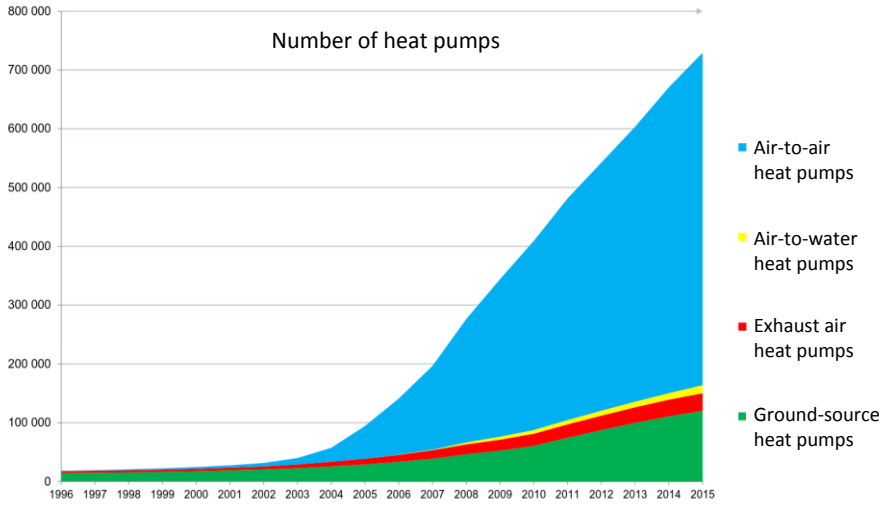


Figure 2.3 Number of heat pumps in Finland. Adapted from (SULPU 2015).

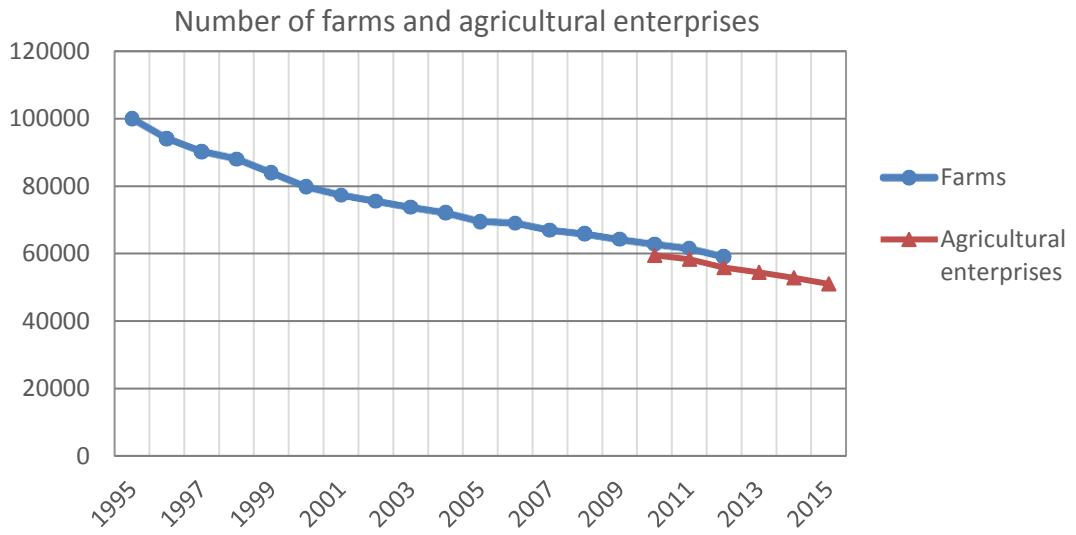


Figure 2.4 Number of farms and agricultural enterprises in Finland. Based on data available in (Luke 2016).

Geographical generalization also causes errors in load profiling. The load profiles are often created to model the average national electricity consumption. They do not take into account the regional differences, which originate from different climate conditions, building stock, and socioeconomic factors. The need to divide a large country into smaller geographical areas when performing load profiling has been acknowledged for example in Dahlström et al. (2011).

The number of customer classes is very low in some cases, for example in the U.K. where only eight customer classes are used, and this can impair the accuracy of load profiling. Outliers, i.e. customers who do not clearly belong to any customer class and whose electricity consumption profile differs from all other customers, are also troublesome for the existing load profiling methods. Especially large outliers are harmful because they can be a source of large (absolute) modelling errors.

In Finland, many DSOs have detected that the existing load profiles are no longer accurate enough and are considering using AMR measurements instead. Some DSOs have already modelled large customers (fuse size ≥3×63 A) with previous year AMR measurements taking into account only the shift in day of the week rhythm. This is not a good approach either and reflects poor trust in the existing load profiles rather than a desirable direction for load profiling. When the previous year AMR measurements are used directly as load models, large errors ensue. The temperatures between years vary considerably and measurements cannot be used as load models without proper temperature correction. Moreover, the measurements do not include estimates for the load variability, which are needed in probabilistic distribution network calculation, or take into account the temporal location of special days.

This thesis aims to fix all the above-mentioned shortcomings in the existing load profiling practices. Improvements and changes are presented in Chapters 3 and 4.

3 Methods for improving load profiling

This chapter presents the author's propositions for fixing the defects in the existing load profiles which were presented in Subsection 2.3.2. Electric load temperature dependency, profile drifts, errors in customer classification, large and exceptionally behaving customers and geographical load diversity are addressed and customer behavior change detection and other possible improvements, which could further increase the load profiling accuracy, are discussed. All the methods presented in this chapter assume that the AMR data has already gone through data validation, where gaps and other gross errors in the measurements have been addressed. Data validation is required from the DSOs and is included also in the upcoming national data hubs (Fingrid 2017a; Statnett 2014).

3.1 Load temperature dependency calculation

It is well known that the weather influences electricity demand in many ways. Outdoor temperature is clearly the most important weather factor, but also solar radiation (day length, time of day, and cloudiness), wind, and humidity affect electricity demand (Meldorf et al. 2007). In this thesis, only the load temperature dependency is taken into account. It has been shown that the outdoor temperature explains the majority of the weather-induced changes in electric load (Siirto 1989). Also, in the used load profile structure, the seasonal variations in day length and daily variations in solar radiation (day and night) are already modelled by the seasonally and hourly varying load expected values. Only the effect of cloudiness is left unmodelled and this defect is partly compensated by the correlation between cloudiness and outdoor temperature. Wind speed and direction can also have some effect on the individual customer's electricity consumption but in general, this effect is very small. According to ASTA II study cited in Siirto (1989), wind increases building heating energy need only by 0.5 % on average.

This thesis uses the following temperature dependency model:

$$\Delta P(t) = \boldsymbol{a}(t) \times (T_{24}(t) - \mathrm{E}[T(t)]), \tag{5}$$

where $\Delta P(t)$ is the outdoor temperature dependent part of the load P at time t, $\boldsymbol{a}(t)$ is the customer class specific load temperature dependency parameter (W/°C), $T_{24}(t)$ is the average outdoor temperature from the previous 24 hours, and $\mathrm{E}[T(t)]$ is the expected value of the outdoor temperature. The expected value $\mathrm{E}[T]$ is a vector containing twelve long term (30 years) monthly average temperatures for the studied location. The average outdoor temperature $T_{24}(t)$ is calculated as:

$$T_{24}(t) = \frac{\sum_{i=t-24}^{t-1} T(i)}{24}, \tag{6}$$

where T is a time series containing hourly average temperatures. The temperature dependency parameter $\boldsymbol{a}$ is a vector containing six values defined separately for each two-

month period starting from January. Also monthly and seasonal (four seasons of the year) temperature dependency parameters were experimented but the monthly parameters were too sensitive to small perturbations in identification data and the seasonal parameters could not model the yearly temperature dynamics as well as the parameters with higher resolution. The two-month division was found to be a good compromise. In the earlier Finnish load research studies, the temperature dependency was defined as a percentage of load change per Celsius degree (Sener 1992; Jalonen et al. 2003). This practice was abandoned in this thesis (although it was still used in [P5]), because it tends to distort the daily load profile shapes by allocating more absolute change to peak load hours than to valley hours. By using a temperature dependency defined as watts per Celsius degrees, the absolute change is the same during all hours of the day regardless of the differences in hourly loadings. This coincides with the way of thinking where heating forms a base load that is independent of the user activity induced load.

The outdoor temperature does not influence the electric load directly but through a delay. Physically, this is caused by the heat stored in the buildings. The length of the delay was studied in Mutanen (2010). Correlations between hourly loads and delayed average temperatures were calculated for each hour of the day with averaging windows of different length. Mean correlation over all the hours of the day was calculated and the window length with the highest mean correlation was chosen as the optimum delay. It was observed that different customer classes have different delays ranging from one hour to over 48 hours. In addition, different hours of the day had different delays. Night hours had long delays and during daytime the delays were shorter. On average, the optimum delay was 24 hours and in the name of simplicity, it was decided that this value is used for all customer classes and all hours of the day. In this thesis, it is thus assumed that the hourly loads depend linearly on the average temperature of the preceding 24 hours.

Strictly speaking, the load temperature dependency is not linear. For example, in summertime, the normally negative temperature dependency can change to positive when the temperature rises and cooling loads increase. In cold countries like Finland, this effect is barely noticeable but in warmer countries, this needs to be taken into account. However, also in Finland the temperature dependency levels off when the temperature rises and eventually ceases to exist. In this thesis, the cut-off temperature was determined experimentally and +19 °C was found to be a suitable limit. In some sources, it is said that in wintertime, the temperature dependency decreases in extremely cold temperature (i.e. below −25 °C) as the heating equipment reach their maximum output (Meldolf et al. 2007). The author has not detected this phenomenon from the AMR data, even the coldest days in the research data reach an average temperature of −32 °C, and this non-linearity is thus not considered in this thesis.

Changes in the load temperature dependency can be observed also when the temperature falls close to zero degrees. This behaviour can be explained with the deployment of additional heaters, for example car engine block and cabin heaters, and is clearly visible in Figure 3.1. This could be modelled by determining different temperature dependency values for different temperature ranges. However, in the chosen model structure, this was not necessary because the adopted two-month division means that the temperature ranges

are already limited by natural temperature variations within each two-month period. In addition, dividing the two-month periods further into different temperature ranges would reduce the size of samples used in temperature dependency determination too much.

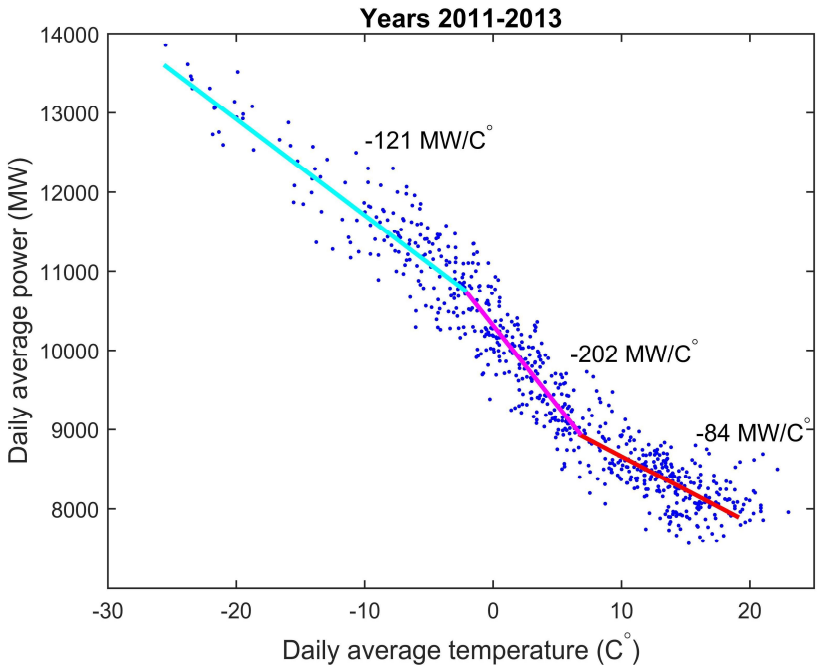


Figure 3.1. Temperature dependency of the total electricity consumption in Finland (workdays only). Based on national consumption data available in Fingrid (2017b).

There are also some specific customer types that have a non-linear temperature dependency. Air-to-air heat pumps have non-linear coefficient of performance (COP) and a minimum operating temperature, and ground-source heat pumps may have been dimensioned to cover only part of the peak heating need. The temperature dependency of a heat pump heated house may therefore be higher than normal in extremely low temperatures. Another special case is houses with storage heaters. When the temperature falls, the power of the storage heater remains constant but the time that the heater is on increases. The modelling of houses with storage heaters has been addressed in literature (Riihimäki & Koponen 2012; Koponen & Niska 2016). The temperature dependency model presented in this thesis is thus not suitable for detailed modelling of all individual customer groups but provides a simple general temperature dependency model for load profiling.

3.1.1 Calculation of temperature dependency parameters

In this thesis, the temperature dependency parameters are determined with linear regression analysis for each two-month period. The effects of daily and monthly fluctuations in electricity demand are eliminated by choosing the dependent and determining variables as follows:

- Dependent variable (regressand): difference between the daily energy consumption and the average daily energy consumption on a similar day (same day of the week and month).

- Independent variable (regressor): difference between the daily average of effective temperatures and the average of effective temperatures on a similar day.

Here the effective temperature means the average temperature of the 24 hours preceding each hour. When the average of effective temperatures over a period of one day is calculated, the result is a weighted average of the hourly temperatures of the previous day and the studied day (excluding the last hour). The hour immediately before the studied day has the highest weight because it affects all the hourly loads in the studied day. Other hours have smaller weights since they affect only some of the hourly loads.

Sometimes the daily energies are so scattered that the temperature dependency parameters cannot be determined reliably. The significance of the relationship between the daily energy and outdoor temperature can be assessed with the correlation coefficient and the Student's t-test. If the correlation is not significant, there is a chance that it is actually zero or opposite in sign than the obtained correlation. The correlation is significant if the value τ, calculated with (7), is larger than the value of τ picked from one tailed t-distribution table with $n-2$ degrees of freedom and a chosen significance level (Lowry 2017).

$$\tau = R\frac{\sqrt{n-2}}{\sqrt{1-R^2}}, \tag{7}$$

where R is the Pearson product-moment correlation coefficient and n is the sample size. In this thesis, when calculating the customer class specific temperature dependency parameters, the significance level is set to 5 %, which is a commonly used limit in statistics. If the significance criterion is not met, a zero temperature dependency is assumed.

3.2 Load profile updating

The electricity consumption habits have evolved over the years and many of the existing customer class load profiles have become outdated. This problem can be corrected by using AMR measurements to update the customer class load profiles. The customer class information for each customer is usually available in NIS and the AMR measurements can be obtained from the meter data management system (MDMS). Now, the update requires only that the AMR measurements are grouped according to the customer classification, summed, and formed into a load profile. The forming should include the calculation of temperature dependency parameters, temperature normalization, calendar correction, and scaling, but overall this would be a rather straightforward process.

In [P4], it was shown that the load profile updating, together with the correct usage of temperature dependency information, can improve the load profiling accuracy by 30 %. In this case, the load profiles were used to perform a day-ahead forecasting of hourly loads for a period of one year, which was not included in the model identification data, and the accuracy was defined as a square sum of the forecasting errors. The studies done in [P6] showed that in all customer classes, the updated load profiles differed clearly from

the original load profiles. Figure 3.2 shows weekly load profiles for four of the most radically changed load profiles. Figure 3.3 compares the measured and modelled loads.

From Figure 3.2 it can be seen that the intra-day load variation in many of the updated load profiles is smaller than the intra-day variation in the original load profiles. While this implies a changed customer behavior, it can also be caused by the deteriorated customer classification. When a customer group becomes more heterogeneous, i.e. it contains differently behaving customers, the overlapping of the load profiles tends to smooth out the daily load profiles. The customer classification errors are addressed in the next section.

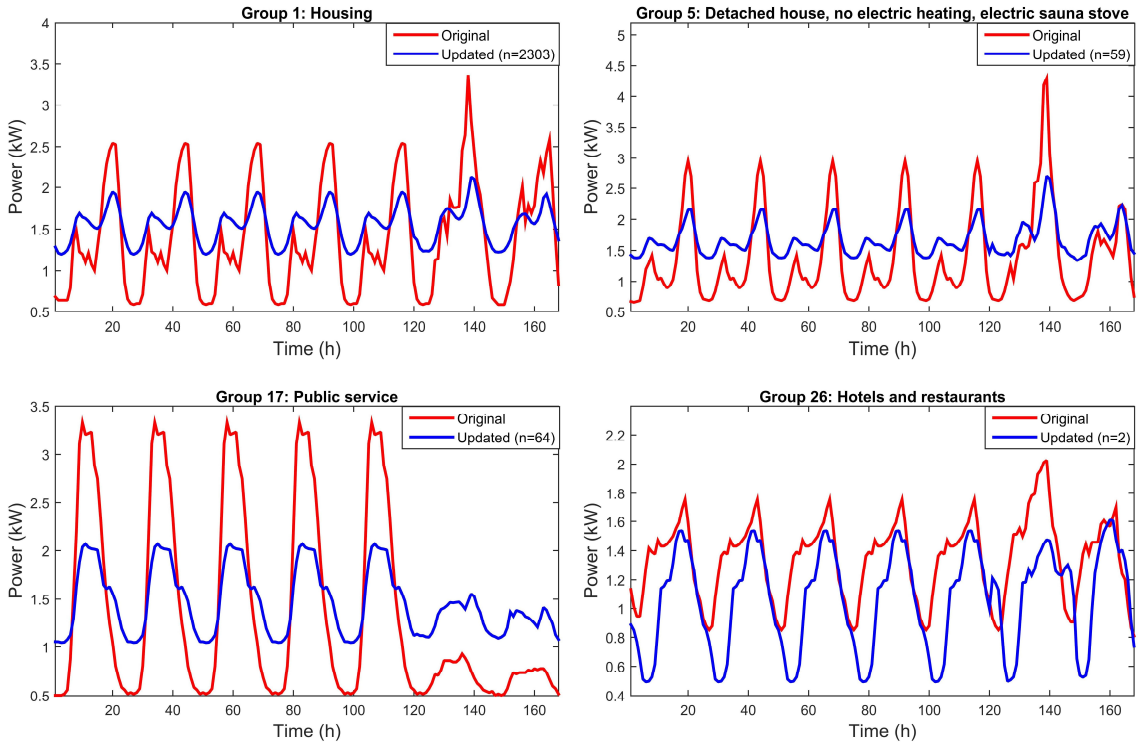


Figure 3.2. Effects of load profile updating. Only a part of the yearly topography, second week of February, is shown in this figure.

3.3 Customer reclassification

As discussed in Subsection 2.3.2, the present customer classification is not up-to-date and classification errors are common. With AMR measurement, the customer classification errors can be corrected. The AMR measurements can be used to determine which existing customer class load profile is closest to each customer's measured load. This can be done, for example, by calculating the Euclidian distance (Han et al. 2012, p. 72) between the measured load and all existing customer class load profiles. The customer class to which the measured load has the smallest distance is then selected as an optimal customer class. While this sounds simple, there are a few issues that need to be taken into account in this comparison. Temperature normalization must be applied to the measured loads so that they represent load at the same long-term average temperature as the load profiles. In addition, calendar correction must be applied so that the days of the week and special days match.

According to the results presented in [P4], the customer reclassification alone improves the load profiling accuracy by 7 %. Further improvements can be achieved if the customer reclassification is combined with the load profile updating. Here lies a pitfall though; if the customer reclassification is done after the load profile updating, the updated load profiles no longer represent the typical behavior of customers classified into that group. To correct this situation, one would need to update the load profiles again, using the new customer classification but after that the customer classification would be again incorrect. For final results, the customer reclassification and load profile updating would have to be iterated until convergence is achieved. The procedure described here is a simple form of clustering and this opens a whole new research topic; use of clustering algorithms in electricity customer classification. Chapter 4 describes in detail how clustering algorithms can be applied to do simultaneous load profile updating and customer reclassification.

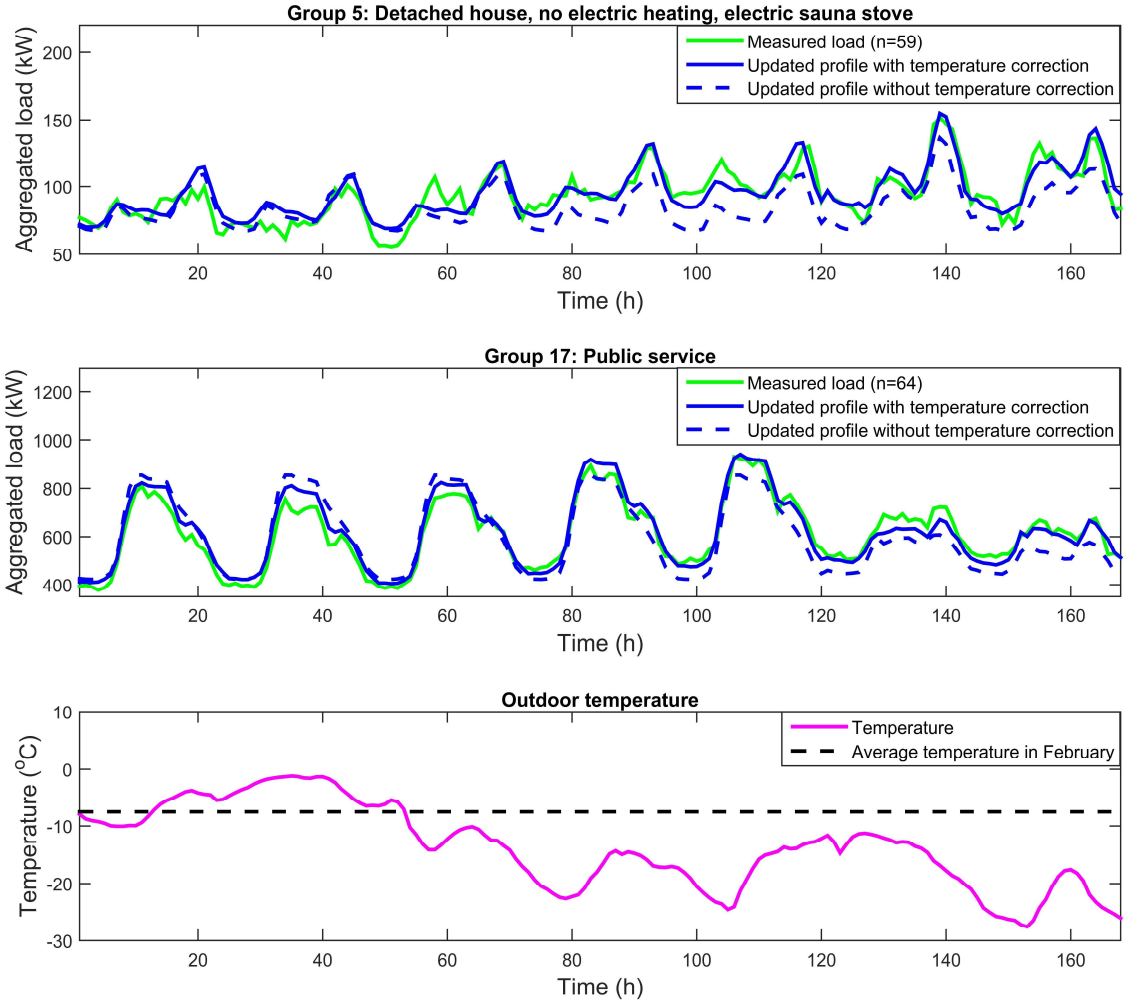


Figure 3.3. Comparison of measured and modelled load during the second week of February 2011.

3.4 Individual load profiles

The previous year AMR measurements should not be used as individual load profiles as such, because they do not take into account the temperature differences between the years or changes in the calendar. Moreover, the measurements do not model the load stochastic variation, which is essential in modern probabilistic distribution network calculation. In

this thesis, a new method for forming individual load profiles is presented. The proposed load profiling method is composed of five steps:

1. Calculation of individual temperature dependency parameters using the method described in Subsection 3.1.1.
2. Temperature normalization so that the measurements correspond to the load in long-term monthly average temperatures.
3. Calculation of type weeks, for both load expected and standard deviation values, during each month.
4. Topography construction for the target year. This uses the above calculated type weeks and takes into account the target year calendar.
5. Scaling to appropriate annual energy, for example to 10 MWh/year standard value or directly to the expected annual energy.

When the individual load profiles are used, temperature correction is applied similarly as with the customer class load profiles. Depending on the application, the individual load profiles are corrected to match either the measured temperature, forecasted temperature, or assumed worst case scenario temperature. The above presented method has some limitations. When measurement data of only one year is available, the monthly type weeks are typically calculated from a sample of four weeks. From a statistical point of view, this is hardly sufficient but provides on average better results than individual load profiles without type weeks, and the results improve if data from several years is available. Figure 3.4 shows how the length of the available AMR data set affects the accuracy of individual load profiles. Results with and without type weeks are also compared. The load temperature dependency and the target year calendar are taken into account in both cases. The only difference is that daily sub-profiles in the type week approach are calculated from a large pool of similar days, whereas in the other case every daily sub-profile has only one similar day per each preceding year.

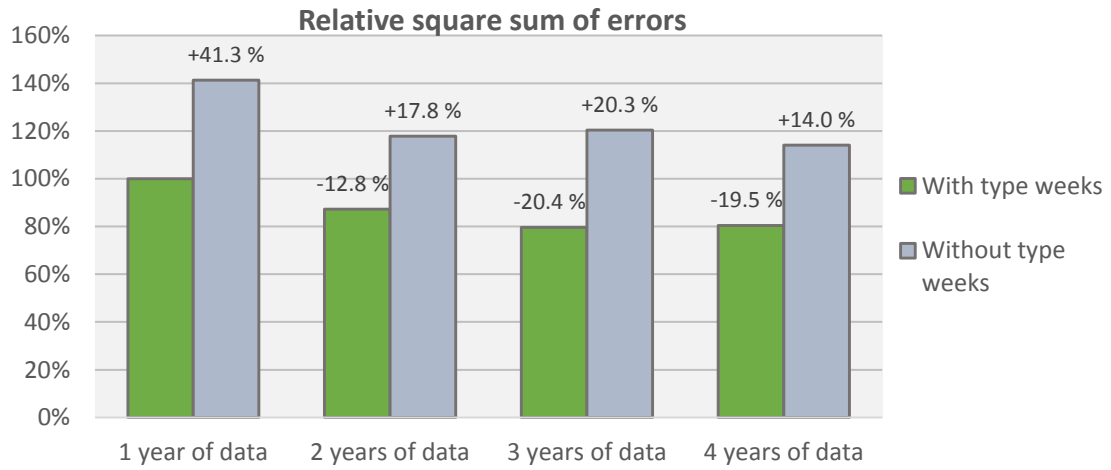


Figure 3.4. Square sum of errors when individual load profiles identified with 1–4 years of AMR data are used to model the consumption of the next year, given in relation to case with type weeks and only one year of identification data.

3.4.1 Comparison with other load profiling methods

In [P4], the individual load profiles were compared with other load profiling methods. When studying small domestic customers, the individual load profiles were clearly better than the existing customer class load profiles but provided only marginal improvement when compared with cluster based load profiles which are later described in Chapter 4. Together with the hugely increased model complexity (the number of load profiles), this result implies that it is not worthwhile to use individual load profiles to model small domestic customers. When larger non-residential customers were studied, the individual load profiles were clearly better than either the existing or cluster based load profiles. The benefit of individual load profiles is undoubtedly larger when they are applied to large customers. Later in Subsection 4.5.2, a method for selecting the customers who benefit the most from individual load profiles is presented.

3.4.2 Improvements to the type weeks

In [P4], each day of the week was modelled separately but in later studies it was discovered that for domestic customers the division into three type days (weekdays, Saturday, and Sunday) is often sufficient. In fact, using only one type day for all weekdays can, in some cases, enhance the load profiling accuracy. This is probably because the sample size increases and the type day for weekdays can be calculated more reliably than the type days for individual weekdays. For large customers, a weekly model with seven distinct type days is often better. Figure 3.5 shows how the electricity consumption of large non-residential customers, which were studied in [P4], varies according to weekday. These large customers exhibit similar behavior on Tuesdays, Wednesdays, and Thursdays but have deviating behavior on Monday mornings and Friday evenings. Even though the differences are small, they are significant because the load stochasticity is low on these customers.

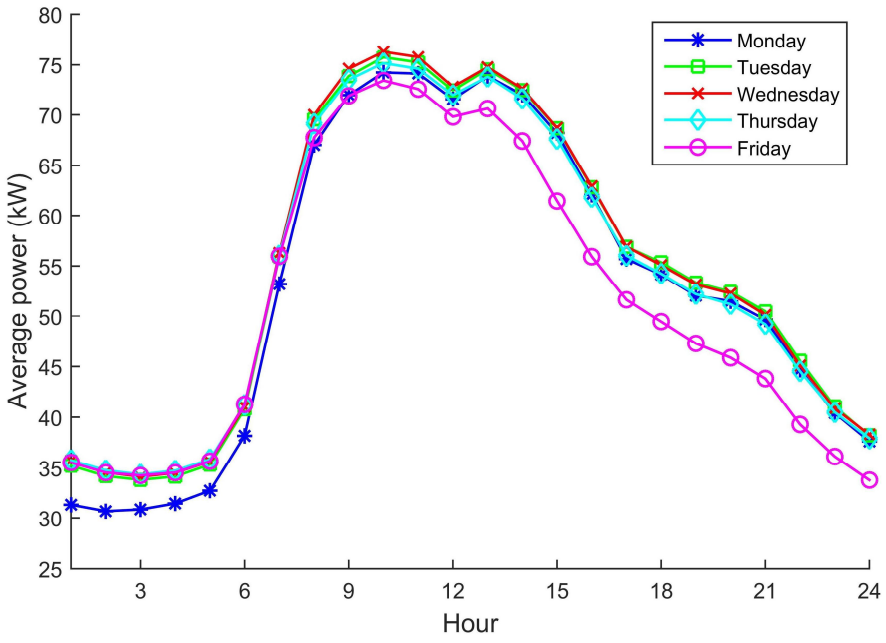


Figure 3.5. Average electricity consumption of large non-residential customers on different weekdays.

In general, small customers are better modelled with three day types and large customers with seven day types. However, this is not always the case and instead of relying on a simple size based division, the author aspired to develop a statistical method for detecting differences in weekday behavior. In [P6] and [P8], one-way analysis of variance (ANOVA) was used to determine if the yearly means of different weekdays are similar or not. One-way ANOVA tests the null hypothesis that the means of three or more independent samples are equal (Lowry 2017). Since the one-way ANOVA examines only one dependent variable, each hour of the day had to be analyzed separately. In this thesis, the weekdays were found to behave differently if the null hypothesis for any hour of the day was rejected. The experiments and visual inspections revealed that with typical excess probability values (e.g. p=0.05), many cases of dissimilar weekdays were left undetected. The excess probability had to be raised up to 35 % in order to detect all the visually identified cases with dissimilar weekdays.

The drawback of using multiple ANOVAs is that it increases type I (false positive) errors. When all 24 hours of the day are analyzed, the probability that at least one null hypothesis is rejected by mere chance is considerably higher than the excess probability used for individual hours. In [P9], this issue was addressed using one-way multivariate analysis of variance (MANOVA). MANOVA is a generalization of ANOVA to a situation in which there are several dependent variables (Tabachnick & Fidell 2006). In this case, the dependent variables are the hourly loads during hours 1–24. When analyzing the differences between weekdays, MANOVA takes into account all hours of the day and their correlations. Matlab function *manova1* (MathWorks 2017) was used to analyze if the yearly means of the weekdays are similar or not. With 5 % excess probability about 30 % of all customers were found to have dissimilar weekdays. This is roughly the same percentage of customers that were found using ANOVA and 35 % excess probability.

After the customers with dissimilar weekdays have been identified, the analysis can be continued with different post hoc procedures. Descriptive discriminant analysis (DDA) for example can be used to analyze which hours contributed the most to the detection of dissimilarity (Warne 2014). In this case, the dissimilarity was most often detected based on the evening hours (18.00–21.00) and Friday was most frequently the day that differed from the other weekdays. In this thesis, the detection of similar weekdays was utilized not only in individual load profiling but also in pattern vector formation, which is part of the developed clustering method presented in Section 4.5.

3.5 Geographically bounded load profiles

The present customer class load profiles usually model electricity consumption on a national level. With AMR measurements, load profiles for more strictly restricted geographical areas can be defined. This will reduce the load profiling errors as the differences in climatic zones, dwelling types, building parameters, and habits in different parts of the country will be taken into account. The differentiating habits can be related, for example, to firewood usage and to activities during the holiday seasons. Also, the sampling errors will vanish if all the customers are metered.

The author of this thesis proposes that each DSO should have their own customer class load profiles. In Finland, there are relatively many distribution network companies; seventy-seven according to Energy Authority (2017). Many of these companies are descendants of the old municipal electricity utilities and operate in only one homogeneous geographical area. In these cases, only one set of load profiles is needed. There are also some larger companies that span over a vast geographical area or own distribution networks in different parts of the country. In those cases, several sets of load profiles are needed to describe the load behavior in different areas. Also, if a company owns network in a large city and in a remote rural area, a distinction between these two areas could be made.

3.6 Customer behavior change detection and other possible improvements

Like the individual load profiles in Figure 3.4, the customer class load profiles become more accurate when they are calculated based on several years of AMR data. This of course applies only if the customer behavior remains constant and does not change over the years. If the customer behavior changes, for example due to heat pump installation or electric vehicle purchase, the pre-change measurements should be discarded and only the post-change data should be used in load profiling. Detecting the changes in customer behavior would facilitate the development of dynamic load profiles that can quickly adapt to changes in the customer behavior. The development of change detection methods is outside the scope of this thesis, but they have been studied by Chen (2014) and Chen et al. (2015).

In this thesis, ANOVA and MANOVA were used to identify whether the weekdays should be modelled with one or five separate type days. The utilization of these methods could be extended to other days of the week, for example to comparisons between Saturdays and Sundays. Also, since weekday dissimilarity was often detected based on only one or two days, only these days could be separated from the other weekdays.

The more detailed modelling of special days could also be one source of improvement. At the moment, the customers are assumed to behave on eves and public holidays similarly as in Saturdays and Sundays, respectively. In reality, the load profiles of certain special days (e.g. Christmas Eve) differs clearly from normal Saturdays and Sundays. Also, it was observed that the weekdays between Christmas and Epiphany differ from typical weekdays in December and January. Especially schools exhibit atypical behavior during this period. MANOVA could again be used to detect which customers behave abnormally during these days.

In future, load profiling needs to take into account several new development trends affecting the electricity end use. These include the growing number of grid-connected microgeneration, introduction of demand response, novel tariff structures, battery energy storages, and smart home automation systems. The approach of this thesis has been to analyze existing AMR data and develop load profiles for the present loads, and therefore some emerging trends, which do not show in the data, have been left with lesser attention.

Depending on which of these trends become significant, they should be taken into account in load profiling. The future development needs of the presented load profiling method are discussed in Section 6.1.

4 Clustering of electricity customers

Clustering is a data analysis technique aimed to determine how the data is organized. Clustering algorithms divide a set of observations into subsets (clusters) so that the observations in the same cluster are similar and the observations in different clusters are dissimilar. The similarity and dissimilarity are usually quantified with some measure of proximity. The outcome of cluster analysis is typically a partitioning where observation is assigned to a cluster. As a result, not only do we know which observations are similar but can also characterize members of each cluster with a cluster centroid. This enables data compression as multiple observations can be summarized with centroids.

Cluster analysis is used in many fields of science, for example biology, medicine, computer science, marketing, finance, and engineering. In the field of electricity distribution, there is often a need to cluster electricity customers into similarly behaving groups. The clustering can be done based on the measured consumption profiles, or quantities calculated from the measurements, and there are many clustering algorithms to choose from.

In this chapter, various ways to perform the electricity customer clustering are presented and discussed. As in earlier chapters, the focus is on electricity *customers* (i.e. electricity users that are metered and billed individually) but the same clustering methods could also be applied for larger electricity consumers consisting of multiple customers (e.g. blocks of flats).

4.1 Clustering algorithms

Thousands of clustering algorithms have been presented in the literature and new ones appear continuously (Jain 2010). For electricity customer clustering alone, dozens of different algorithms have been applied or proposed. For example; iterative refinement clustering (Batrinu et al. 2005), hierarchical clustering (Chicco et al. 2005), fuzzy *c*-means clustering (Lo et al. 2005), modified follow-the-leader clustering (Carpaneto et al. 2006), support vector clustering (Chicco & Ilie 2009), *k*-means clustering (Räsänen & Kolehmainen 2009), ant colony clustering (Chicco et al. 2013), subspace projection based clustering (Piao et al. 2014), multi-resolution clustering (Li et al. 2016), and spectral clustering (Vercamer et al. 2016).

There are so many clustering algorithms that the literature has found it useful to categorize them. Han et al. (2012) classifies clustering algorithms into four main categories; partitioning methods, hierarchical methods, density-based methods, and grid-based methods. In this section, some clustering algorithms belonging to these main categories are presented. Many other classifications also exist, for example, classification to hard and soft (fuzzy) clustering methods. In hard clustering an object belongs to exactly one cluster, while in soft clustering the object belongs to each cluster with a certain degree of membership. Moreover, the categorization itself can sometimes be fuzzy.

4.1.1 Partitioning methods

Given a set of n objects, the partitional clustering methods construct k partitions of the data, where each partition represents a cluster, $k \leq n$, and each cluster contains at least one object. Most partitioning methods are distance-based. After the number of partitions (k) and the initial partitioning have been defined, *iterative relocation technique* is used to improve the partitioning by moving objects from one group to another. In general, the objects are assigned to the closest or the most similar cluster. Several different distance metrics can be used, although the Euclidean distance is the most common. When the Euclidian distance is used, the partitional methods find spherical clusters. (Jain 2010; Han et al. 2012)

Achieving global optimality in partitioning-based clustering is often computationally prohibitive, potentially requiring an exhaustive search where all possible partitions are tested. To overcome this problem, greedy approaches are often used in practice. A prime example of a greedy approach is the *k*-means algorithm, which progressively improves the clustering solution and approaches a local optimum. (Han et al. 2012)

***K*-means** is still one of the most widely used clustering algorithms, even though it was first introduced over 60 years ago. Easy implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity (Jain 2010). The credit for inventing the k-means algorithm is usually given either to Lloyd (proposed in 1957, published in 1982), Forgy (1965, cited in Bock 2007), or MacQueen (1967) who was the first to use the term "*k*-means". However, algorithms with similar principles have been presented even earlier, for example by Steinhaus (1956, cited in Bock 2007).

K-means is a centroid-based partitioning technique, meaning that the clusters are represented by central vectors called *centroids*. The centroids can be defined in various ways, such as by the mean or medoid of the objects assigned to the cluster. The goal of the *k*-means algorithm is to minimize the sum of squared distances between all objects and their assigned clusters. The objective function J_k to minimize is:

$$J_k = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} dist(\boldsymbol{x}, \boldsymbol{c}_i)^2, \tag{8}$$

where $\boldsymbol{x}$ is the point in space representing a given object, $\boldsymbol{c_i}$ is the centroid of the cluster C_i (both $\boldsymbol{x}$ and $\boldsymbol{c_i}$ are multidimensional), k is the number of clusters, and *dist* is the chosen distance metric (usually the Euclidean distance). This objective function aims to make the clusters as compact and separate as possible. The *k*-means algorithm is summarized in Algorithm 4.1. The inputs to the *k*-means algorithm are the number of clusters (k) and a data set containing n objects. The outputs are partitioning of these n objects into k clusters and cluster centroids. (Han et al. 2012)

One of the disadvantages of the *k*-means is that the number of clusters needs to be defined a priori. Selecting the right number of clusters is not a trivial task. In this thesis, the selection of the optimal number of clusters, from the load profiling point of view, is studied in Section 4.4. Another drawback of the *k*-means is that it converges only to local minima and different initializations can lead to different results. One way to mitigate this

issue is to run the *k*-means algorithm several times with different initial partitions and choose only the result that gives the smallest objective function value (8). Another way is to improve the initial partitioning, for example with sub-sample clustering. Several other methods for improving the initial partitioning have also been proposed in the literature, for example, the *k*-means++ (Arthur & Vassilvitskii 2007) and scalable *k*-means++ (Bahmani et al. 2012) methods.

Algorithm 4.1. The *k*-means procedure (Han et al. 2012).

1: Randomly choose *k* objects from the data set as the initial cluster centroids (output: $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\ldots,\boldsymbol{c}_k\}$).
2: Set *idx = 0;* *#Initialize cluster indices*
3: Set *repeat = 1;* *#Initialize loop condition*
4: **while** *repeat*
5: (Re)assign each object to the cluster to whose centroid ($\boldsymbol{c}_i$) the object has the shortest distance (output: idx_{new})
6: Update the centroids (output: $\boldsymbol{C}$)
7: **if** $idx_{new} == idx$
8: Set *repeat = 0;* *#Terminate loop*
9: **end**
10: Set $idx = idx_{new}$; *#Update cluster indices*
11: **end**

ISODATA (iterative self-organizing data analysis technique) is another early clustering algorithm originally proposed by (Ball & Hall 1965). The basic principle of ISODATA is similar to that of *k*-means, but additional steps to split heterogeneous clusters and merge neighboring clusters have been added. The algorithm is described in detail in [P5] where it is used to cluster electricity customers. It is often claimed that the ISODATA algorithm can find the number of clusters automatically, but in reality a sensible initial guess $k_{initial}$ is required and the final number of clusters is usually within range $[k_{initial}/2,\ 2\times k_{initial}]$. Also, the user-given splitting and merging thresholds affect the final number of clusters as is shown in [P5].

Fuzzy *c*-means is also very similar to the *k*-means. The main difference is that in the fuzzy *c*-means algorithm the objects are not forced to belong to only one cluster. Instead, they are assigned membership degrees between zero and one, which enables them to belong to several clusters. The fuzzy *c*-means algorithm was first proposed by Dunn (1973) and later improved by Bezdek (1981, cited in Jain 2010). The goal of the algorithm is to minimize the weighted sum of squared distances between all objects and their assigned clusters. The objective function J_c to be minimized is:

$$J_c = \sum_{i=1}^{k}\sum_{j=1}^{n} \boldsymbol{W}_{i,j}^{m} \times dist(\boldsymbol{x}, \boldsymbol{c}_i)^2, \tag{9}$$

where $\boldsymbol{W}$ is a weight matrix and m is a parameter that determines the influence of the weights (Liao 2005). Similarly as the *k*-means, the fuzzy *c*-means procedure requires the number of clusters as an input and starts from a random partitioning. The initial partitioning is then improved iteratively. The cluster centroids are calculated as weighted means and the weight matrix is updated during every iteration.

Gaussian mixture model (GMM) clustering is another soft clustering method closely related to the *k*-means method. In GMM clustering, the clusters are modelled with

Gaussian probability density functions (PDFs) and individual objects are allowed to have memberships to several clusters. The whole data set is therefore modelled by a mixture of Gaussian components. For a multivariate case, the joint PDF of the mixture model is:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \, N(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), \tag{10}$$

where k is the number of mixture components (clusters), w_i is the weight of the ith component, $\boldsymbol{\mu}_i$ is the mean of the component, and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the component (Singh et al. 2010). The goal of GMM is to define the set of parameters $\boldsymbol{\theta} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{k}$ so that the log-likelihood of (10) is maximized. This cannot be done analytically, and therefore the expectation-maximization (EM) algorithm is usually applied to find a local maximum (Ari et al. 2012). Like the *k*-means, the EM algorithm starts from a random initial estimate of parameters ($\boldsymbol{\theta}$) and improves this iteratively; E-step computes the weights and M-step computes the Gaussian parameters. Unlike in *k*-means, in GMM the clusters can have non-spherical (elliptical) shapes. If hard memberships are used instead of soft memberships and identity covariance matrices are assumed, the GMM yields similar results as the *k*-means.

Mixtures of factor analyzers (MFA) clustering reshapes GMM by applying factor analysis to reduce the number of parameters in the component-covariance matrix of (10). Factor analysis, which is a statistical method for modelling the covariance structure of high dimensional data using a small number of latent variables, is extended to a mixture model that allows different local factors in different regions of the input space. This results in a model which concurrently performs clustering and dimension reduction, and is essentially a reduced dimension mixture of Gaussians. The MFA model is given by (10), where the ith component-covariance matrix has the form:

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i, \tag{11}$$

where $\boldsymbol{B}_i$ is a $q \times d$ matrix of factor loadings and $\mathbf{D}_i$ is a diagonal matrix. Here, q is the dimensionality of the original data, and d is the number of subspace dimensions. The parameters $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$, along with weights w_i and means $\boldsymbol{\mu}_i$, can be determined using the EM algorithm or some variation of it. (Ghahramani & Hinton 1997; McLachlan et al. 2003)

4.1.2 Hierarchical methods

By default, the hierarchical clustering methods do not provide a single partitioning of the data set. Instead, they give an extensive hierarchy of clusters that merge with each other at certain distances. The most natural way to represent this hierarchy is through a tree-shaped structure called *dendrogram*. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all the groups are merged into one, or a termination condition is reached. The divisive approach, also called the top-down approach, starts with all the objects in

the same cluster. In each successive iteration, a cluster is split into smaller clusters, until each object is in a cluster of its own, or a termination condition is reached. (Han et al. 2012)

Agglomerative methods are more popular than the divisive methods because splitting large clusters into smaller ones is more challenging than merging small clusters. There are $2^{n-1}–1$ possible ways to split a set of n objects into two subsets. If n is large, it is computationally prohibitive to examine all the possibilities and usually some type of heuristic method for splitting is used. Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This leads to small computation costs but may cause inaccurate partitioning, since the erroneous decisions cannot be corrected. (Han et al. 2012)

The properties of hierarchical clustering depend very much on the choice of the used distance function and the linkage criterion. The most common linkage criteria are: single linkage, complete linkage, average linkage, centroid method, and Ward's method. An algorithm using single linkage measures the minimum distance between the clusters and is sometimes called a nearest-neighbor algorithm. A complete linkage algorithm measures the maximum distance between the clusters and is sometimes called a farthest-neighbor algorithm. The average linkage is calculated as a mean over all pairwise object distances between two clusters and the centroid method calculates the distance between the cluster centroids. In Ward's method, the distance between two clusters is defined as an increase in the total intra-cluster sum of squared errors when the clusters are merged. (Rencher 2002; Han et al. 2012)

Single and complete linkage algorithms are sensitive to outliers and noisy data. The single linkage algorithm suffers also from the so-called chaining phenomenon, where consecutive merges can lead to a situation where clusters at the ends of the chain are very distant to each other. Complete linkage algorithm on the other hand favors equally sized clusters, which can be good or bad depending on the structure of the data. The use of average linkage and centroid method alleviates the outlier sensitivity problem and provides a compromise between minimum and maximum distances. (Han et al. 2012)

Several variations of the basic algorithm have been proposed in the literature. For example, the BIRCH (balanced iterative reducing and clustering using hierarchies) algorithm proposed by Zhang et al. (1996), the Chameleon algorithm proposed by Karypis et al. (1999), and the CURE (clustering using representatives) algorithm proposed by Guha et al. (2001).

4.1.3 Density-based methods

Partitioning and hierarchical methods have difficulties in finding arbitrarily shaped clusters. To find arbitrarily shaped clusters, one can model clusters as dense regions in the data space, separated by sparse regions. This is the main idea behind density-based clustering methods. One of the most popular density based clustering methods is DBSCAN (density based spatial clustering of applications with noise). DBSCAN scans the data and finds core objects that have dense neighborhoods. A user-defined parameter

Eps is used to specify the radius of the neighborhood and the neighborhood is determined to be dense, if the number of objects within the neighborhood is greater than or equal to a user-specified parameter *MinPts*. A cluster is formed by a group of connected core objects and all other objects that are reachable (within *Eps*) from these core objects. (Ester et al. 1996; Han et al. 2012)

As with many other clustering algorithms, the downside of DBSCAN is that it requires user-defined parameters which are often difficult to choose and tune. The later variants of DBSCAN have addressed this issue, for example, the DENCLUE (density based clustering) method proposed by Hinneburg and Keim (1998), and the OPTICS (ordering points to identify the clustering structure) method proposed by Ankerst et al. (1999).

4.1.4 Grid-based methods

Grid-based clustering methods quantize the object space into a finite number of cells that form a grid structure. The clustering is then performed on the grid, instead of the original object space. This reduces the processing time since the number of grid cells is typically smaller than the number of objects in the original space. The downside is that the clustering accuracy is limited by the granularity of the grid. (Han et al. 2012)

The grid-based clustering algorithms are good in clustering very large data sets. STING (statistical information grid) and Wave Cluster algorithms, for example, can efficiently cluster large spatial data sets. Their computational complexity is linearly proportional to the number of cells at the lowest grid level (STING) or to the number of objects (Wave Cluster). CLIQUE (clustering in quest) and MAFIA (merging of adaptive intervals approach to spatial data mining) are examples of grid-based algorithms suitable for clustering numerical data. They scale well in relation to the number of objects, but their time complexity is exponential in the number of dimensions. (Ilango & Mohan 2010)

4.2 Comparison of clustering methods

Scientists often search for the best method for solving a certain problem, but finding it is not always possible. For example in this case, it is impossible to determine the best algorithm for electricity customer clustering. First of all, one lifetime is not enough to compare all the clustering algorithms. Countless, but not all, clustering algorithms are suited for electricity customer clustering. Some clustering algorithms do not work well with electricity consumption data due to the high dimensionality of the data. Secondly, since there is no universal definition for a *cluster*, the countless proposed cluster validity indices weight the cluster properties differently. This diversity of cluster validity indices is evident in Subsection 4.4.1 where several validity indices are used to determine the optimum number of clusters. Often in practice, researchers select a validity index that coincides with their subjective idea of a cluster, and thus human bias is incorporated into the results. In his position paper, Estivill-Castro (2002) wisely argues that clusters are in the eye of the beholder, and that is the reason why so many cluster validity indices and subsequently also clustering algorithms have been proposed.

It is difficult, if not impossible, to find the best method for electricity customer clustering but this has not stopped the efforts. Gerbec et al. (2003a) compared hierarchical and fuzzy

c-means clustering in load profile classification and ended up recommending fuzzy *c*-means over the hierarchical clustering. However, in their other paper Gerbec et al. (2003b), they only state that both the fuzzy *c*-means and the hierarchical clustering with Ward linkage criterion yield similar results.

Chicco et al. (2005) compared *k*-means, fuzzy *c*-means, self-organizing map (SOM), modified follow-the-leader procedure, and agglomerative hierarchical clustering with both Ward and average linkage criteria. The hierarchical clustering run with the average linkage criterion and the modified follow-the-leader algorithm were found to be the two most effective algorithms for clustering daily load profiles.

Tsekouras et al. (2007) compared modified *k*-means, fuzzy *c*-means, adaptive vector quantization, and hierarchical clustering with seven different linkage criteria. According to three cluster validity indices, the modified *k*-means was the best, according to two validity indices the adaptive vector quantization was the best, and according to one validity index the hierarchical clustering with Ward linkage criterion was the best.

Kim et al. (2011) compared *k*-means, fuzzy *c*-means, and hierarchical clustering. The *k*-means algorithm was found to be the most accurate one when clustering daily load profiles.

Chicco et al. (2012) compared *k*-means, fuzzy *c*-means, follow-the-leader algorithm, and hierarchical clustering with six different linkage criteria. The result was that the *k*-means algorithm was the fastest but the hierarchical clustering with single linkage criterion was the best according to the cluster validity indices. However, the inspection of hierarchical clustering results revealed that the majority of the clusters were comprised of outliers and the bulk of daily load profiles was concentrated in only one cluster. The *k*-means algorithm, on the other hand, created many uniformly sized clusters; as is desirable.

As this short literature review shows, there is no clear consensus which clustering method is the best for electricity customer clustering. Although it should be noted that the k-means method won two out of the four comparisons it participated in and was, in this author's opinion, a moral winner in the third even though the other methods achieved better cluster validity index values (q.v. previous paragraph). The *k*-means method should therefore provide a safe starting point for clustering electricity customers. It is also the default clustering method in the two-stage clustering method developed in this thesis. Several other clustering methods were also tested during the development and comparisons are presented in Subsection 4.5.3.

4.3 Dimension reduction

Electricity customer clustering is often done based on high-dimensional time series data. In literature, the most common approach is to cluster the customers based on daily consumption data that has 24, 48, or 96 dimensions, depending on whether the measurements are done hourly, half-hourly, or quarter-hourly. It is possible to perform the clustering based on this raw data but dimension reduction is often applied to speed up the clustering, reduce noise in the input data, and mitigate the effects of the curse of dimensionality. If the electricity consumption of an entire year is considered as a whole,

as is done in this thesis, the number of dimensions is measured in thousands. With this many dimensions, the need for dimension reduction becomes even more pronounced.

There are several different dimension reduction methods and many of them have been applied to electricity customer clustering. Räsänen and Kolehmainen (2009) divided the hourly consumption data into weekly windows and extracted features describing the weekly mean, standard deviation, skewness, kurtosis, chaos, energy, and periodicity. The clustering was then done based on these feature vectors, which were considerably shorter than the original time series. Verdú et al. (2004) used nine features describing the shape of the daily load pattern, for example, the ration of average daytime load to maximum daytime load and the ratio of average daytime load to average night time load. Good results were achieved also when the 96 dimensional daily load patterns were transformed into 24 dimensional hourly load profiles. The feature extraction can also be done in the frequency domain, as has been done by Verdú et al. (2004) and Carpaneto et al. (2006). They have used discrete Fourier transform to compute the amplitude and phase of the harmonics present in the daily load patterns. In Mets et al. (2016), fast wavelet transformation was used to represent the 96 dimensional daily load patterns with only seven features.

When clustering time series data longer than one day, representative load patterns (RLPs) are often used to reduce the input data dimensionality. The RLP can be either a single typical daily profile (TDP) or a vector of TDPs describing the average load on different days of the week and seasons or months. The vector approach is used for example by Dang-Ha et al. (2016). Although, a more popular approach is to use single TDPs and perform the clustering separately for each loading condition (day of the week and season) as has been proposed, for example, by Chicco et al. (2013).

One of the most used dimension reduction techniques is the principal component analysis (PCA). In PCA, the goal is to reduce the dimensionality of the data set while retaining as much information as possible. In mathematical terms, PCA performs an orthogonal linear projection of high dimensional data onto a low dimensional subspace so that the variance of the projection is maximized. The greatest variance lies on the first subspace dimension (principal component) and each following dimension, which are orthogonal to the previous dimensions, explains as much as possible of the remaining variability. Usually, a significant amount of variance present in the data can be explained with a number of principal components that is only a fraction of the original number of dimensions. PCA has been used in numerous publications to reduce the dimensionality of the electricity consumption data prior to clustering (Cheng & Li 2009; Koivisto et al. 2013; Lu et al. 2016).

Self-organizing maps (a.k.a. Kohonen networks) are also used often in dimension reduction. SOM is a type of artificial neural network that is trained to project the input space into a reduced dimension space (usually into a two-dimensional hexagonal map), where the proximity properties of the input space are approximately preserved. In general, the SOM may be considered as a nonlinear generalization of PCA (Dang-Ha et al. 2016). There are many publications where SOM has been applied to electricity consumption data

(Figueiredo et al. 2005; Räsänen et al. 2010; Niska 2013). It is also possible to chain dimension reduction techniques, for example, load profile shape characterization can be followed by SOM.

4.4 Selecting the optimal number of clusters

A major challenge in cluster analysis is the selection of the "right" number of clusters. The number of clusters is a crucial input parameter in many clustering algorithms, such as *k*-means, fuzzy *c*-means and BIRCH. Hierarchical clustering methods do not require the number of clusters as inputs, but the operator must decide where to cut the hierarchical tree into clusters and this is an analogous problem to selecting the number of clusters. Some clustering algorithms, such as DBSCAN and OPTICS, determine the number of clusters automatically but they require other input parameters that are equally difficult to optimize. This section discusses how to find the optimal number of clusters for a k-means algorithm applied to electricity customer classification.

4.4.1 Cluster validity indices

In electricity customer classification, the true customer classes are unknown and therefore external evaluation cannot be used to assess the classification accuracy. Instead, internal evaluation based on the clustered data must be used. Internal evaluation methods usually give the best score to the algorithm that produces clusters with high intra-cluster similarity and low inter-cluster similarity. However, different cluster validity indices weight these attributes differently and the results vary. A clustering algorithm that aims to minimize a certain criterion and uses a certain distance metric, naturally gets a good score from an evaluation method that uses similar objective function and distance metric. For example, the k-means algorithm performs well when evaluated with the sum of squared errors (SSE) and compared to other clustering algorithms.

The variability of the results is not limited to the selection of the best clustering method. Even if the clustering method has already been selected, the different cluster validity indices give different results for the optimal number of clusters. This is true even with relatively simple low-dimensional data sets as has been shown in (Baarsch & Celebi 2012), (Wu & Yang 2005) and (Tibshirani et al. 2001). In this thesis, clustering is done on high-dimensional electricity consumption data and it is very unlikely that all the cluster validity indices perform well with this data set. However, reliable validity indices are needed when determining the optimal number of clusters. Next, nine different cluster validity indices are tested and analyzed, and the one most applicable to this problem is selected. The tested validity indices are:

1) Davies–Bouldin index (DBI) (Davies & Bouldin 1979)
2) Dunn index (Dunn 1973)
3) Silhouette (Rousseeuw 1987)
4) Mean index adequacy (MIA) (Chicco et al. 2003)
5) Clustering dispersion indicator (CDI) (Chicco et al. 2003)
6) Calinski–Harabasz Criterion (CH) (Kryszczuk & Hurley 2010)
7) Bayesian information criterion (BIC) (Schwarz 1978)

8) Akaike information criterion (AIC) (Akaike 1974)
9) Sum of squared errors (SSE) (Duda et al. 2012, p. 542).

In this test case, the goal is to find an optimal number of clusters for a standard *k*-means method which is used to cluster pattern vectors of 6425 electricity customers. In this case, the pattern vectors consist of 864 values describing the average weekly consumption (working day, Saturday, and Sunday) on 12 different months. The calculations are done with random initialization, ten replicates and squared Euclidian distance as a distance metric. Figure 4.1 shows the index values as a function of *k* and Table 4.1 displays the optimal number of clusters. Only nine internal validity indices were compared here, although tens of others also exist, for example scatter index (Pitt & Kirschen 1999), gap statistics (Tibshirani et al. 2001), ration of within cluster sum of squares to between cluster variation (Tsekouras et al. 2007), partition coefficient and exponential separation index (Wu & Yang 2005), Bezdek's partition coefficient (Wu & Yang 2005), Xie-Beni index (Xie & Beni 1991), and WB-index (Zhao & Fränti 2009).

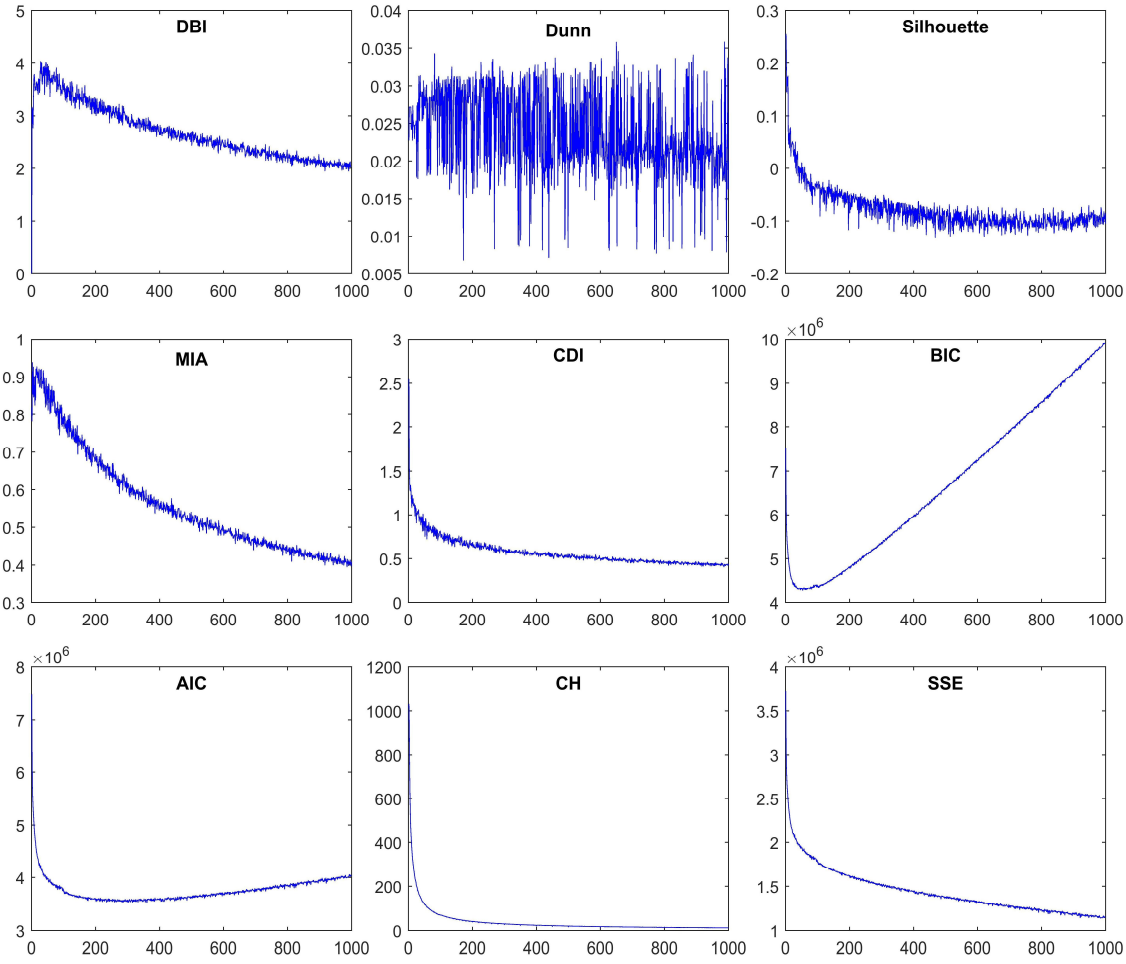


Figure 4.1. Cluster validity index values as a function of *k*.

From Figure 4.1 and Table 4.1 it can be seen that several of the tested cluster validity indices do not provide usable results with the electricity usage pattern data. The Dunn index is so volatile that it is useless. The Silhouette and CH reach their maximums when *k*=2. This is clearly too small a value for this application. The MIA, CDI and DBI do not

reach their minimum values with a reasonable number of clusters (minimums at k>1000). Moreover, MIA, CDI, DBI, Dunn and Silhouette appear to be very sensitive to small random changes in clustering, which happen every time the k-means algorithm is run. Figure 4.2 shows how DBI and SSE vary when they are applied to results from different k-means runs. The DBI values vary a lot and it is impossible to get reliable results with a single run. SSE on the other hand provides consistent results on every run.

Table 4.1. Optimal number of clusters based on the studied cluster validity indices.

	Optimal number of clusters	Type of optimum point	Clearness of the plot
DBI	>1000	Minimum	Poor
Dunn	988	Maximum	Useless
Silhouette	2	Maximum	Poor
MIA	>1000	Minimum	Poor
CDI	>1000	Minimum	Poor
BIC	50–70*	Minimum	Good
AIC	200–400*	Minimum	Adequate
CH	2	Maximum	Excellent
SSE	60–80*	Knee	Good

* Only approximate values are given due to the volatility of the curve or subjective nature of the optimum point.

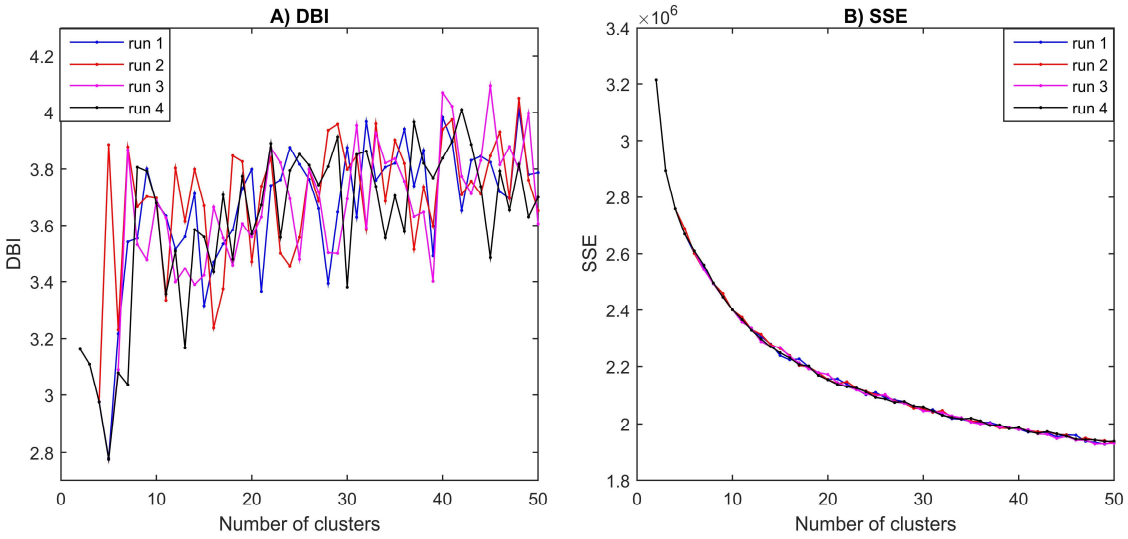


Figure 4.2. Comparison of cluster validity index volatility.

BIC and AIC provide minimum points where the optimum number of clusters should be. However, the solutions provided by the BIC and AIC are very different. The different results are caused by the different penalty terms used in BIC and AIC. BIC penalises model complexity more than AIC (when sample size n>7). In the case of k-means algorithm, the BIC and AIC are basically knee point detection methods where the knee point is found from the location on SSE curve that has the same slope as the penalty term.

It is possible to determine the optimal number of clusters directly from the SSE curve using also other knee point detection methods, which are studied in the next section.

4.4.2 Knee point detection

Locating the "knee" of an error curve in order to determine the optimal number of clusters is a well-known method. With human intuition it is easy to see the knee region in a figure. However, defining a universal application-independent knee point detection algorithm that works with all kinds of knee curves is not as easy as one might think.

The knee of a curve is best defined as the point of maximum curvature. Curvature is a mathematical measure of how much a function differs from a straight line. As a result, maximum curvature captures the levelling off effect used to identify knees. For any continuous function *f*, there is a standard closed-form $K_f(x)$ that defines the curvature of *f* at any point as a function of its first and second derivatives (Satopää et al. 2006):

$$K_f(x) = \frac{f''(x)}{(1 + f'(x)^2)^{3/2}} \tag{12}$$

While curvature is well-defined for continuous functions, it is not easy to apply for discrete data sets, such as the curves studied in this thesis. In a discrete case, it would be possible to determine the curvature by fitting a continuous function on the data. However, fitting a continuous function to a set of arbitrary data points is difficult, especially if the data is noisy.

The other knee point detection methods presented in literature can be divided into local and global methods. Local methods are based on geometric features calculated using information from only a few neighboring points on a curve. Examples of such methods are the angle-based methods presented in (Dep & Gupta 2010) and (Branke et al. 2004), and the Menger curvature based method described in (Satopää et al. 2006). The local knee point detection methods do not work well with noisy data and are therefore not suitable for solving the knee point detection problem in this thesis. Although Figure 4.2 showed that SSE is a lot more stable index than DBI, there is still enough noise to render local knee point detection methods useless. The small variations present in the studied SSE curve are highlighted in Figure 4.3.

Global knee point detection methods aim to take the overall trend of a curve into consideration when determining the knee point. The most well-known global method is the Normal-Boundary Intersection (NBI) method where a straight line is drawn from the first point of the curve to the last point of the curve and a knee point is declared at a point that is farthest from this line (Das 1999). Euclidian distance is usually used to determine the distance from the line but also vertical distance can be used, as has been done in the Kneedle algorithm (Satopää et al. 2006). The curve start and end points can also be used to define angles between lines going through each point of the curve and the aforementioned start and end points (Dep & Gupta 2011). The knee point is at the point that has the smallest (or largest, depends on the formulation) bend-angle. The L-method (Salvador & Chan 2004) fits straight lines to the left and right sides of each point and selects the point which has the smallest sum of weighted root mean squared errors of the

fitted lines. In situations where the curve has a long tail (more points after the knee than before it), the L-method tends to give too large values for the knee. Salvador and Chan (2004) have presented a modified L-method to suppress this problem.

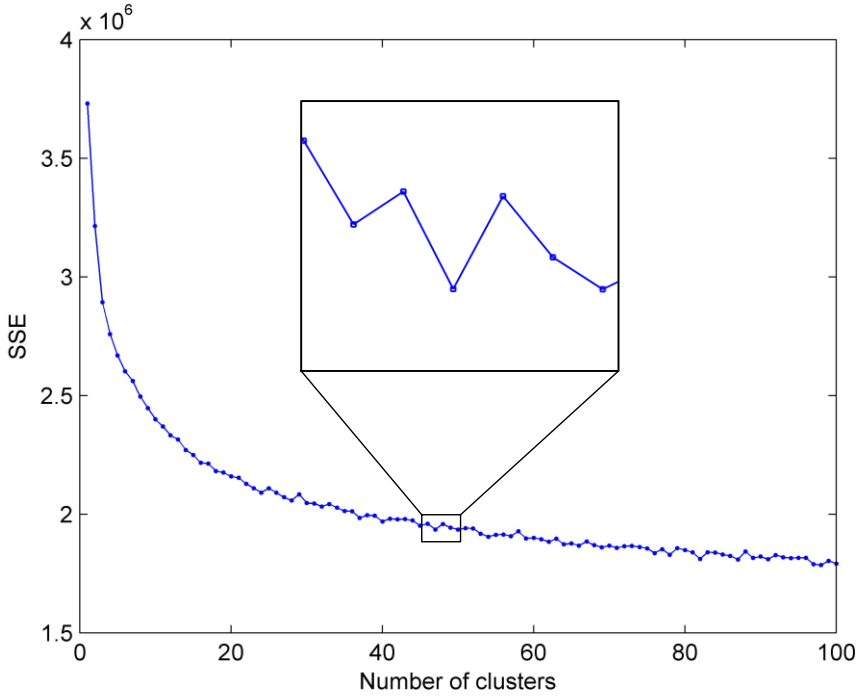


Figure 4.3. Small variation in SSE that render local knee point detection methods useless.

The above-mentioned global knee point detection methods were applied to the previously calculated SSE curve. Table 4.2 shows the results. There are clear differences between the methods. Moreover, the results are affected by the range selected for inspection. If only one hundred first SSE values are studied, the knee point is found between 10 and 18 clusters. If the whole range from one to 6425 is studied, the knee point is found between 6 and 769 clusters.

Table 4.2. The knee point location found with different knee point detection methods and with different input ranges.

	Range 1–100	**Range 1–1000**	**Range 1–6425**
Fitted function ($y = a \cdot x^b + c$) + curvature calculation	14	79	7 (bad fit)
NBI	16	82	450
Kneedle	16	82	450
Bend-angle	16	54	140
L-method	18	176	769
Modified L-method	10	10	6

Only NBI and Kneedle provided similar results with all the studied ranges. The bend-angle method gave smaller values when the range was wide and the curvature calculation for full range was unreliable due to bad fit. The L-method gave very large values and the modified L-method gave very small values. Moreover, the disparity between the knee

point detection methods is not the only issue. If *k*-means clustering method is used, this kind of knee point detection requires that the clustering is repeated with all viable values of *k* and this is very compute-intensive.

Since the knee point detection has turned out to be a challenging task and the optimum number of clusters cannot be determined unambiguously, in this thesis, the final number of clusters is selected based on other criteria such as the intelligibility of the model. Hundreds of clusters (i.e. customer classes) could be too much for the DSO staff to handle, and reducing the number of customer classes from the present level would feel like a step backward. In addition, new customer classes are needed to capture the effects of new emerging technologies, such as electric vehicles, home automation systems, and micro generation. From this perspective, a number moderately larger than the present number of customer classes would be ideal. In [P4], [P6], and [P9] the accuracy of cluster profiles was compared with Sener profiles and the number of clusters was chosen to be the same as the number of existing customer classes. This way, the effect of the number of the customer classes was eliminated from the comparison.

4.5 The developed load profiling procedure

The load profiling procedure developed in this thesis is shown in Figure 4.4. The procedure contains load profile updating, individual load profiling and a two-stage clustering method. The load profile updating is included in the load profiling procedure, because this research was started at a time when the AMR roll-out in Finland was not yet completed. The use of individual load profiles and cluster profiles requires that AMR measurements are available. If the AMR measurements are missing, the updated load profiles are used as a backup. Nowadays, the AMR systems cover almost 100 % of the customers, but there are still situations when individual or cluster profiles cannot be used. For example, when a new house is built and connected to the network, electricity consumption history does not exists and the individual load profile cannot be formed nor the customer can be classified based on the consumption history.

It is rather straightforward to use updated load profiles, individual load profiles and cluster profiles side by side. This was demonstrated with a modified prototype version of the ABB MicroSCADA Pro DMS 600 –software. The DMS 600 is connected to a database that contains a customer information table. One column of this table contains the original customer classification information. The load profiles to which this classification refers to were updated and two additional columns were added to the customer information table; one column for the cluster information and one column for the individual load profile numbers. A minor modification was made to the DMS 600 so that the program reads first the column with individual load profile numbers, continues to read the cluster information only if the individual number is missing, and finally proceeds to read the original customer class if the cluster information is missing. In other words, the individual load profiles were prioritized over cluster profiles, which in turn were prioritized over the updated load profiles. For testing purposes, the DMS 600 prototype allowed the user to choose whether or not the individual load profiles and the cluster profiles were used in the network calculation.

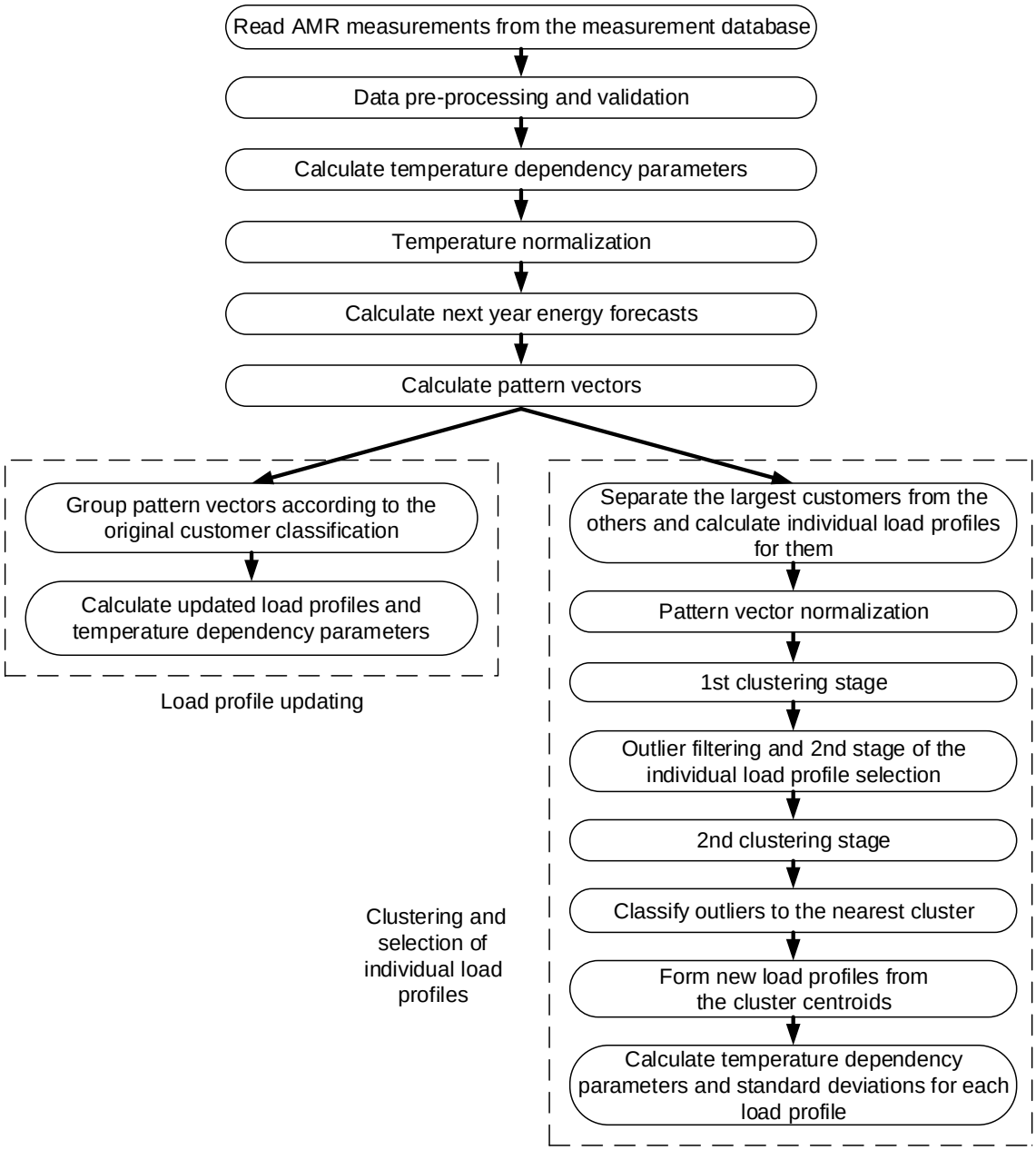


Figure 4.4. Flow chart for load profile updating and clustering.

The load profiling procedure shown in Figure 4.4 has been implemented as a Matlab program. It starts by importing the AMR and temperature measurements to Matlab and continues with data pre-processing and validation. Small gaps in the data are interpolated, exceptionally large or small values are labeled as bad data, and the data format is checked (e.g. the unit of the data, cumulative or non-cumulative time series). Then the temperature dependency parameters are calculated individually for each customer using the method presented in Subsection 3.1.1 and the time series are normalized to correspond to consumption in the long-term monthly average temperatures. This allows us to treat the measurements equally, even if they are originally from different years with different temperatures.

The next year energy consumption forecasts for each customer are calculated based on the temperature normalized measurements. If measurement data is available from several

years and there is a statistically significant linear trend in customer level yearly energy consumption, linear extrapolation is used to forecast the next year's energy consumption. The yearly energy forecasts are based on the temperature normalized measurement history and will therefore reflect the yearly energy consumed during a year with average monthly temperatures. Yearly energy forecasts are needed because in network calculation applications the load profiles are scaled to match the expected yearly energy consumption. In addition, the yearly energies are used later in the two-stage clustering procedure.

Pattern vectors describing the average hourly consumption of each day of the week in each month are calculated from the temperature normalized electricity consumption time series. The statistical methods presented in Subsection 3.4.2 are then used to analyze whether or not the weekdays should be modelled separately or with a common weekday model. Either way, the resulting pattern vectors consist of $24\times7\times12=2016$ elements. Some of the customers with similar weekdays could be modelled with pattern vectors consisting of $24\times3\times12=864$ elements, but since we need to compare them with customers who have dissimilar weekdays, all pattern vectors must be of equal length. In case of similar weekdays, the identified weekday model is simply repeated five times. After this, either load profile updating or cluster analysis is performed. These separate procedures have been described in the next two subsections.

4.5.1 Load profile updating

The load profile updating is a simple process. First the pattern vectors are grouped according to the original customer classification, and then the averages of pattern vectors in each group are calculated. These averages are used to calculate the updated load profiles which are formed by extending the pattern vector averages to cover the whole target year and by normalizing the yearly energy consumptions to a standard value of 10 MWh/year. Finally, the customer class temperature dependency parameters are determined using the temperature measurements and means of AMR measurements belonging to each customer class. In here, means are used instead of sums because they are less sensitive to missing data. For the same reason, the pattern vectors are used in load profile updating instead of the temperature normalized measurements.

4.5.2 Two-stage clustering

The proposed clustering procedure starts with the separation of large customers. The largest customers (measured by yearly energy) are separated from the others and assigned for individual load profiling. This is done so that the largest customers do not distort the first stage clustering results. In addition, these large customers would very likely be selected for individual load profiling in later stages anyway. Next, the pattern vectors are normalized so that all vectors have a mean value of one. The previously calculated yearly energies are later used as weights that offset the effect the normalization has on the cluster means.

By default, the first clustering stage uses a weighted *k*-means algorithm developed by the author. In weighted *k*-means, the calculation of distances from cluster centroids is done as in *k*-means, but the normalized pattern vectors are weighted with the corresponding

yearly energies when the centroids are updated. The clustering is initialized with the existing customer classification. Also other initialization and clustering methods could be used. The effect of different initialization and clustering methods on the accuracy of the developed clustering procedure has been analysed in Subsection 4.5.3.

After the first clustering stage, outlier filtering and the second stage of the individual load profile selection are performed. The customers with the largest unweighted distances from the cluster centroids are labelled as outliers and set aside. Empirically, it was observed that removing approximately 10 % of the total population as outliers was sufficient. The customers with the largest weighted distances from the nearest cluster centroids are selected for individual profiling. Figure 4.5 shows the outlier filtering limit and the large individuals that were selected already before the first clustering stage. Most of the customers labelled as outliers have very small yearly energies. Figure 4.6 shows the limit which is used to select the rest of the customers for individual load profiling. Here, the customers with large weighted distances from the closest cluster centroid (*minD*×*E*) are denoted as weighty individuals. Both the outlier percentage and the number of individual load profiles are user-selectable parameters.

The second clustering stage repeats the weighted *k*-means clustering. This time without the outliers and customers that have been selected for individual profiling. The first-stage clustering results are used to initialize the second-stage clustering. After the two-stage clustering, the previously removed outliers are classified to the nearest cluster. Only full pattern vectors are used in the clustering and the incomplete pattern vector (i.e. vectors with gaps) are classified in this stage to the cluster with the most similar load profile shape.

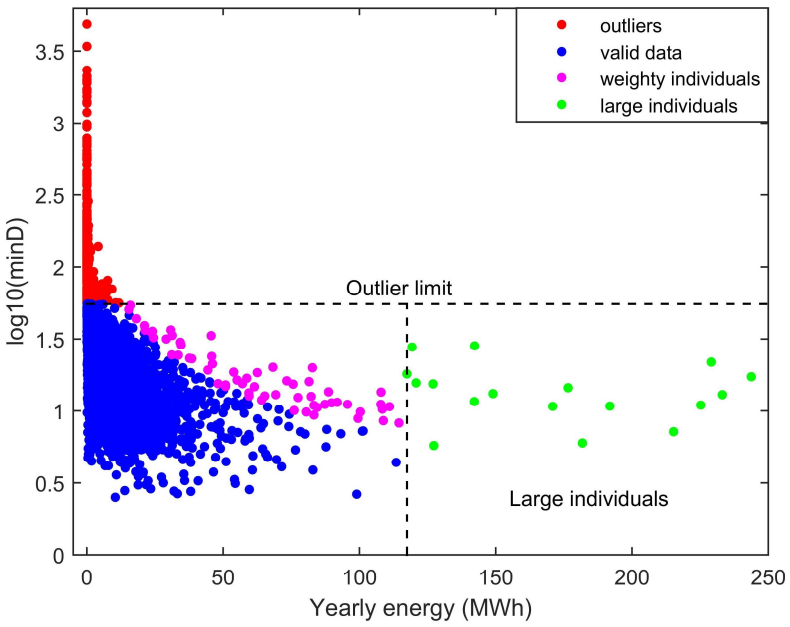


Figure 4.5. Outlier filtering based on the minimum distance from the closest centroid (*minD*) and selection of customer for individual load profiling based on the yearly energy consumption.

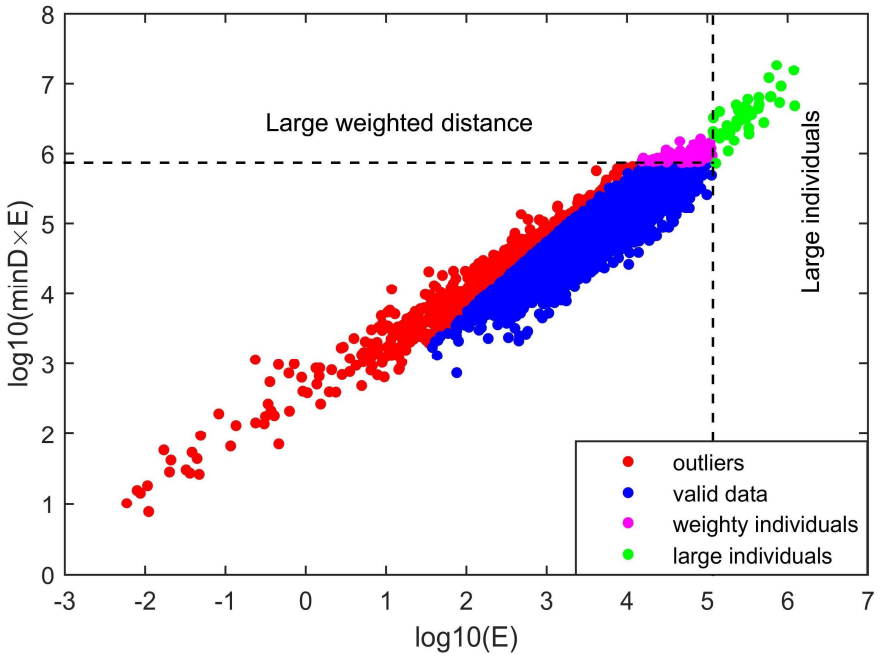


Figure 4.6. Selection of customers for individual load profiling based on the largest weighted distance from the closest cluster centroid. In this figure, *E* is the yearly energy in kilowatt-hours.

The final load profiles are formed from the cluster centroids by extending them to cover the whole target year. Calendar information on the target year is needed in this step. The temperature dependency parameters and standard deviations for each load profile are calculated using the AMR measurements (temperature parameters), temperature normalized AMR measurements (standard deviations), temperature measurements, and classification obtained from the second clustering stage. Both the cluster and individual profiles are made compatible with the existing load profile format where each hour of the year has an expected value and a standard deviation.

4.5.3 Sensitivity to initialization and clustering method

The *k*-means clustering algorithm is very sensitive to initialization, i.e. the final accuracy depends on the objects that are randomly selected as cluster centroids before the first iteration round. The randomness in initialization explains why the *k*-means algorithm often provides different results on different runs. The sensitivity of the proposed two-stage clustering method is studied in Figure 4.7, which shows how the load profiling performance changes when the *k*-means initialization method is varied. The figure is based on the AMR data used in [P9] and the performance was measured by evaluating how accurately the load profiles produced by the two-stage clustering method model the aggregated load of 7532 customers. The model accuracy, on a separate verification year, was measured with mean absolute percentage error (MAPE). In this study, the two-stage clustering algorithm was run 100 times (when applicable), the number of customer classes was set to 37, and individual load profiles were not used.

By default, the proposed two-stage clustering algorithm uses the original customer classification available in CIS as a starting point and in this case the weighted *k*-means algorithm always converges to the same result. With random initialization, the results

were on average poorer but occasionally better results were achieved. The use of subsample clustering did not improve the accuracy when compared with random initialization.

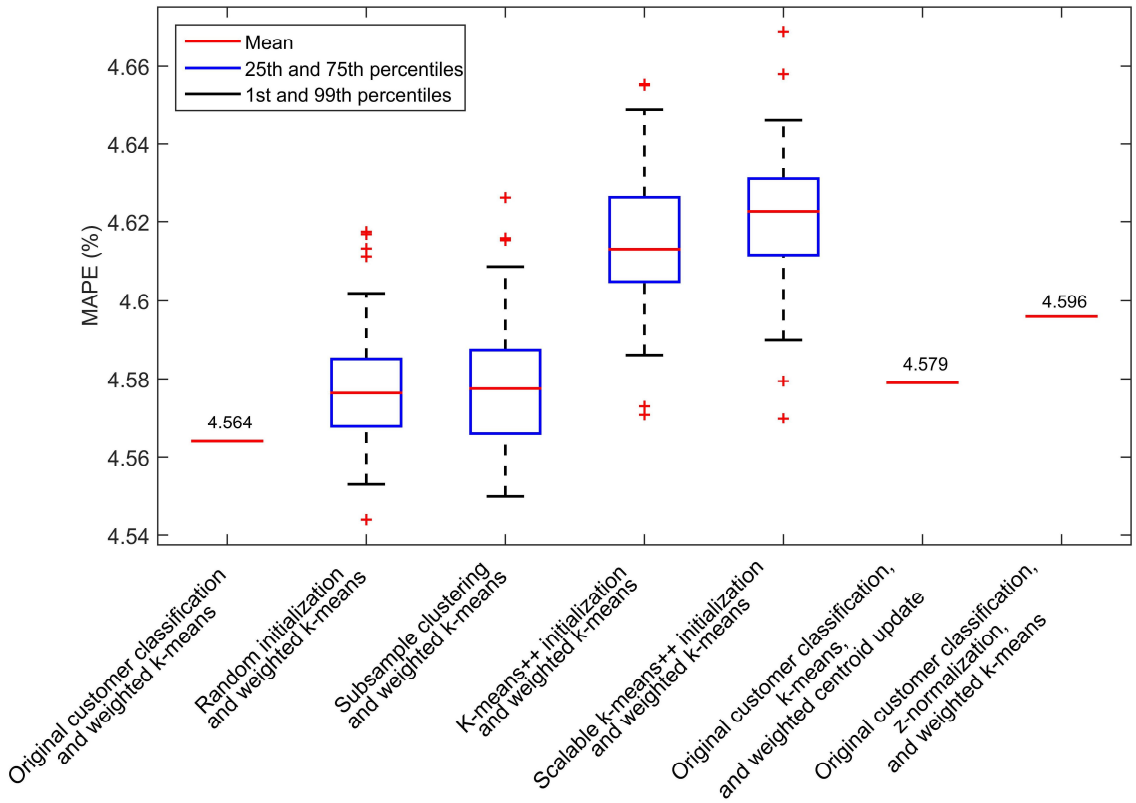


Figure 4.7. The effect of initialization, *k*-means version, and normalization on the accuracy of the developed two-stage clustering algorithm.

In this case, the initialization methods used in *k*-means++ and scalable *k*-means++ algorithms did not improve the results. In fact, they made the results worse. During the initialization phase, these algorithms favor objects that are far away from the already selected cluster centroids. When the data set contains outliers, the outliers are more likely to be selected as initial cluster centroids, because they are far away from all other objects. When using these algorithms, the outlier filtering should be done before the clustering. These two algorithms are thus not suitable for being used with the proposed two-stage clustering method.

In addition to the initialization method, the selected clustering algorithm and the data normalization method also have an effect on the final accuracy. When the classical *k*-means algorithm was used instead of the developed weighted *k*-means algorithm, the accuracy was slightly poorer. However, it should be noted that weighted centroid update must be performed after the classical *k*-means, otherwise the MAPE will be a whole one percent unit higher. It is a common approach to perform standard score normalization (a.k.a. z-normalization) on the data prior to clustering. However, in this case normalization to zero mean and variance of one did not improve the results. On the contrary, the results got worse.

The possibility to replace the *k*-means algorithm with other clustering algorithms was also studied. Figure 4.8 shows the resulting accuracy when the *k*-means algorithm in the proposed two stage clustering method was replaced with different clustering algorithms. The first four algorithms in Figure 4.8 were initialized with the original customer classification. Where indicated, weighted versions of algorithms were used (i.e. the size of the customer was taken into account as a weighting factor in cluster centroid calculation). The last four algorithms used random initialization and the clustering was run 100 times.

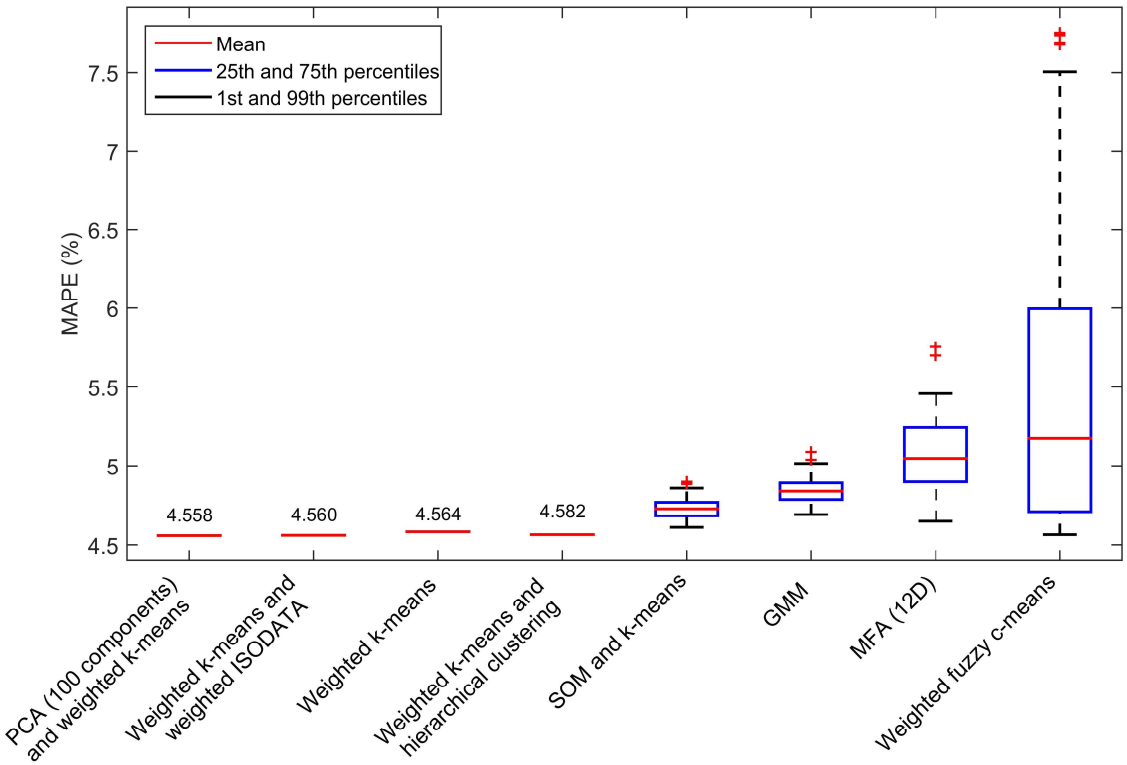


Figure 4.8. The effect of clustering algorithm choice on the accuracy of the developed two-stage clustering method.

The best results were achieved with a combination of PCA and *k*-means. First the input data (pattern vectors) dimension was reduced with PCA and then the weighted k-means algorithm was used for clustering. This was repeated for both the first and second clustering stage. The results shown here were attained by using the 100 first principal components, which in this case explained 85 % of the pattern vector variance. The differences in accuracy were negligible between the four best clustering algorithms but this combination had the shortest overall execution time. The drawback of PCA is that it requires more memory than *k*-means and this can become a limiting factor when clustering very large datasets (>200 000 customers).

The second best results were achieved by using *k*-means in the first clustering stage and ISODATA in the second clustering stage. However, when comparing with the algorithm that uses *k*-means on both stages, the improvement was marginal and there was a 63-fold increase in the second stage execution time. Average execution times for each tested clustering algorithm are given in Table 4.3. The reported execution times were achieved with a desktop computer with Intel Core i7-2600 processor and 16 GB of RAM memory.

Table 4.3. Average execution times for algorithms shown in Figure 4.8 (sorted in ascending order according to total execution time).

	Dimension reduction (1st+2nd stage)	Clustering (1st+2nd stage)	Total
PCA and weighted *k*-means	8+7	23+30	68
Weighted *k*-means	-	87+29	116
Weighted *k*-means and hierarchical clustering	-	87+52	139
SOM and *k*-means	95+89	2+2	188
Weighted fuzzy *c*-means	-	465+58	523
MFA (12D)	-	584+131	715
GMM	-	624+743	1367
Weighted *k*-means and weighted ISODATA	-	87+1827	1914

Using *k*-means in the first clustering stage and hierarchical clustering (with Euclidian distance and Ward's method) in the second clustering stage also provided good results. Hierarchical clustering could not be used in the first clustering stage because it yields poor results in the presence of outliers. The second tested dimension reduction method, SOM, did not perform as well as PCA. Both the final clustering accuracy and execution time were poorer. Different grid sizes were tested and the best results were achieved with a 15×15 hexagonal grid.

In [P7], GMM and MFA were successfully used to cluster 48 dimensional daily load profiles. However, with the 2016 dimensional pattern vectors used here, the clustering accuracy and execution time were clearly inferior to the previously mentioned clustering methods. The fuzzy *c*-means algorithm also performed badly with the studied high dimensional data. With a typically used blending parameter (m=2), many of the cluster centroids coincided and the memberships became approximately equal. Acceptable results were achieved only with very small blending values. This is a well-known problem and several revised fuzzy *c*-means algorithms have been proposed in the literature, for example in (Di Nuovo & Catania 2008; Winkler et al. 2012). The results here were achieved with blending parameter m=1.05.

Although not shown in the results, DBSCAN was also tested but it struggled to find enough clusters. It was able to separate only the most distinct customer types (street lighting, industrial customers, and others). If DBSCAN did find more clusters, they were typically very small and the majority of the customers were clustered into one big cluster.

5 Distribution system state estimation

The purpose of distribution system state estimation (DSSE) is to obtain the best possible estimate of the network state by processing the available information. Usually, the network state means node voltages, line power flows and line current flows. The available information used in DSSE includes network topology, network configuration, line parameters, measurements and load profiles. Traditionally, DSSE relies mainly on primary substation measurements and load profiles. The substation measurements include real-time measurements of busbar voltages and feeder current or power flows. With these measurements, it is possible to estimate the feeder total loads accurately, but the load distributions inside the feeders remain uncertain.

The advent of smart grids has changed the network operation principles and increased the amount of real-time measurements. New measurements are installed along the MV network, to secondary substations and to customer connection points. These measurements not only improve the MV network state estimation accuracy but also enable, for the first time, real-time LV network state estimation. The smart metering infrastructure can be used to improve the state estimation accuracy either by reading the meters in real-time or by using the data collected from customer level electricity usage to improve the load profiles that are commonly used as pseudo-measurements in state estimation. This chapter reviews the available state estimation methods, presents the developed state estimator, and combines the previously presented AMR-based load profiles with the DSSE.

5.1 Literature review

In order to utilise all the new measurements, new state estimation methods are needed. During the past 20 years, countless new DSSE methods have been proposed in the literature. Many of them are based on the weighted least squares (WLS) method but the selection of state variables varies. Some are using node voltages as state variables whereas others have chosen to use branch currents (q.v. Subsection 5.1.1). In addition, several other types of state estimators have been suggested.

5.1.1 Weighted least squares estimation

The objective of state estimation is to determine the most likely state of the system based on the quantities that are measured. One way to accomplish this is by the maximum likelihood estimation, a method widely used in statistics. If the measurement errors are assumed to be normally distributed, the likelihood maximization corresponds to minimizing the weighted sum of squares of the measurement residuals. The weighting of measurements depends on the measurement accuracy. Accurate measurements have large weights and inaccurate measurements have small weights. (Abur & Expósito 2004)

If the network topology and parameters are perfectly known, the network state can be defined, for example, with node voltage magnitudes and angles or with branch current magnitudes and angles. In state estimation, these variables are called *state variables* and all other measurable network variables: node voltages, loads, line power flows and line current flows can be defined as a function of these variables. In literature, the selection of state variables varies. Some are using node voltages whereas others have chosen to use branch currents.

The **basic WLS formulation** is fixed regardless of the chosen state variables. The most likely network state is the one that minimizes the weighted differences between measured network variables and their estimated values. This can be expressed as a minimization problem:

$$\min_{\boldsymbol{y}} J(\boldsymbol{y}) = min_{\boldsymbol{y}} \sum_{i=1}^{N_m} \frac{[z_i - h_i(\boldsymbol{y})]^2}{\sigma_i^2}, \tag{13}$$

where $J(\boldsymbol{y})$ is the objective function to be minimized
$\boldsymbol{y}$ is the state vector that contains all state variables
z_i is value of measurement *i*
$h_i(\boldsymbol{y})$ is measured variable *i* as a function of the state variables
σ_i^2 is variance of measurement *i*
N_m is number of measurements.

If measurements and measurement functions are presented in a vector form and measurement variances are presented in a matrix form, (13) can be expressed as:

$$\min_{\boldsymbol{y}} J(\boldsymbol{y}) = [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{y})]^T \boldsymbol{R}^{-1} [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{y})], \tag{14}$$

where $\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{N_m} \end{bmatrix}$ (measurement vector)

$\boldsymbol{h}(\boldsymbol{y}) = \begin{bmatrix} h_1(\boldsymbol{y}) \\ h_2(\boldsymbol{y}) \\ \vdots \\ h_{N_m}(\boldsymbol{y}) \end{bmatrix}$ (measurement functions)

$\boldsymbol{R} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N_m}^2 \end{bmatrix}$ (covariance matrix).

The minimum of cost function $J(\boldsymbol{y})$ can be found by differentiating it and searching for the zero point. The cost function derivative in respect to state vector **y** is equal to its gradient. Therefore, the state vector minimizing the cost function forces the gradient to zero. The gradient of $J(\boldsymbol{y})$ is:

$$\nabla J(\boldsymbol{y}) = -2\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{z} + 2\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H} \boldsymbol{y}, \tag{15}$$

where $\boldsymbol{H} = \left[\frac{\partial \boldsymbol{h}(\boldsymbol{y})}{\partial \boldsymbol{y}}\right]$ (Jacobian matrix).

When the gradient is zero, we can solve $\boldsymbol{y}$ from (15):

$$\boldsymbol{y} = (\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{z} \tag{16}$$

Since (16) is non-linear, solving the state vector $\boldsymbol{y}$ requires the use of iterative methods, such as the Newton-Raphson method. On every iteration round, a linearized approximation of the state vector change $\Delta \boldsymbol{y}$, shown in (17), is added to the initial state vector value. The iteration is continued until $\Delta \boldsymbol{y}$ is smaller than the predefined threshold. (Abur & Expósito 2004)

$$\Delta \boldsymbol{y} = (\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{y})] \tag{17}$$

Node voltage based estimation algorithms have been used in transmission system state estimation since the 1970s. To speed up the calculation, the traditional transmission system state estimators use *fast decoupled state estimation* where the dependencies between active power and voltage magnitude and reactive power and voltage angle have been eliminated. The fast decoupled method assumes that line resistances are substantially smaller than line reactances. This assumption is not valid for distribution networks and the decoupling cannot be used to speed up DSSE. (Abur & Expósito 2004)

Another common assumption that is made to speed up the calculation is that the Jacobian matrix stays constant during the iteration. This assumption is invalid if the network contains current measurements that are very common in distribution networks. Moreover, current measurements can cause multiple possible solutions and slow down the convergence of the state estimation algorithm. In transmission networks, current measurements can be handled as supplementary measurements since the measurement redundancy is high and amount of current measurements is small. In distribution networks, measurement redundancy is low and it is important to fully utilize all available measurements, including the current measurements.

Despite the above-mentioned problems, the transmission system state estimation principle has been successfully applied to distribution systems in many studies, for example in (Baran & Kelley 1994; Lu et al. 1995; Lin & Teng 1996; Wan & Miu 2003; Cobelo et al. 2007). Work has also been done to improve the current measurement handling capabilities and computational speed (Baran & Kelley 1994; Handschin et al. 1995; Lu et al. 1995; Lin & Teng 1996). In distribution networks, there are many nodes with zero loads and zero production. The load and production on these nodes can be forced to zero by using virtual measurements with very high weights. However, the combination of virtual measurements and pseudo-measurement, which have very low weights, can lead to ill-conditioning of the gain matrix. Equality constraints have been introduced to the WLS formulation to solve this problem (Lin & Teng 1996; Abur & Expósito 2004).

Branch current based estimation algorithm was developed since the voltage based state estimators have problems with the current measurements needed in DSSE. Compared with the node voltage based methods, the branch current method has several

benefits: it is faster, it is not affected by the line R/X-ratio, current magnitude measurements are easier to use with it, and equations are simpler. Moreover, the new method handles power measurements efficiently. This is important since load pseudo-measurements have a vital role in DSSE. (Baran & Kelley 1995; Lin et al. 2001; Teng 2002)

The original branch current based state estimation method developed by Baran and Kelley (1995) has some defects. It cannot handle voltage measurements and is able to calculate only weakly meshed networks. In later publications, the branch current method has been improved. The calculation speed has been further enhanced (Lin et al. 2001) and the ability to use voltage measurements has been added (Teng 2002; Wang & Schulz 2004). Additionally, it has been proposed that current magnitudes and angles could be used as state variables instead of real and imaginary current components (Wang & Schulz 2004). The benefit of using current magnitudes and angles is that there is no need to make an initial guess for the current angle, instead it is automatically estimated based on the (pseudo-)measurements. Also, current magnitude measurements correspond directly to state variables and this simplifies equations. Capability to utilize phasor measurement units (PMUs) is added to branch current based DSSE in (Pau et al. 2013).

5.1.2 Other DSSE methods

Several non-conventional methods have been proposed for solving the DSSE problem. The variety of the proposed methods is wide but they all aim to utilize the available information efficiently and address some of the shortcomings in the previously presented WLS methods.

Fuzzy logic based DSSE algorithms have been developed in (Sarić & Ćirić 2003) and (Pereira et al. 2004). These state estimators incorporate information affected by uncertainty by using fuzzy set theory. For example, historical data can be used to derive typical load profiles defining a band of possible values. Using these typical load profiles, it is possible to obtain fuzzy assessments for active and reactive loads. Furthermore, one can obtain fuzzy assessments as a translation of natural language propositions from experienced operators. Typically, they have a lot of qualitative information expressed in a non-mathematical way. These expressions from human language are transformed into fuzzy numbers and used as fuzzy measurements.

A **hybrid particle swarm optimization** for distribution system state estimation has been proposed in (Naka et al. 2003). Conventional WLS methods assume that the objective functions to be minimized are differentiable and continuous. However, certain equipment in distribution systems have non-linear characteristics and this causes non-linearity to the objective functions. Particle swarm optimization (PSO) can be applied to non-linear and non-continuous optimization problems. A hybrid PSO adds an evolutionary selection mechanism to PSO and can generate high-quality solutions. Another hybrid method based on the combination of Nelder-Mead simplex search and particle swarm optimization is proposed in (Niknam & Firouzi 2009). Although the hybrid PSO is shown to be more efficient than the other evolutionary optimization algorithms, the execution times

reported by Nanchian et al. (2017) indicate that it is not fast enough for real-time DSSE applications.

The use of **neural networks** in DSSE was first proposed by Bernieri et al. (1996) and later several others have studied this approach (Ferdowsi et al. 2014; Barbeiro et al. 2015; Pertl et al. 2016). The NN-based DSSE methods have several benefits: the network state can be estimated even when the grid topology and parameters are unknown; they are robust (no convergence issues) and computationally light (after training) and thus suitable for real-time monitoring; and if only voltage estimates are needed, power injection measurements or models are not necessary. The downsides are that the NN training is computationally intense and requires historical measurements from all the desired NN outputs, the accuracy is compromised if the network state is outside the range covered in the training data, and the NN needs to be retrained if the network or the customer behavior changes.

A **probabilistic approach** to DSSE is presented in (Ghosh et al. 1997). Ghosh points out that the WLS estimation methods incorrectly assume that all the measurement errors are normally distributed. Since load profiles are used as pseudo-measurements, this assumption implies that also loads are normally distributed. This is not true for distribution network loads. Distribution network loads are actually closer to beta or lognormal distributions than normal distributions. To address this issue, a probabilistic DSSE method that accounts for non-normally distributed loads and incorporates load correlations, is proposed. The method utilizes backward/forward sweep calculation and resembles more probabilistic load flow than state estimation. Although the voltage and current measurements are taken into account when calculating the corresponding probabilities, loads are not corrected to match with the line power or current flow measurements.

In smart grid control, it is beneficial to know not only the estimated network states but also the confidence intervals of the estimated states. In recent years, there has therefore been a newfound interest in probabilistic DSSE. Střelec et al. (2015) have taken a similar backward/forward sweep based approach as Ghosh and calculate state estimate probabilities in a presence of photovoltaic energy sources. Brinkmann and Negnevisky (2016) have used the WLS method and extract the state variable variances directly from the inverted gain matrix.

Interior point optimization has been applied to DSSE in (Džafić et al. 2011). This approach is a combination of load flow based scaling and interior point optimization. Interior point methods are known to be fast and scalable (Gondzio 2012) and Džafić shows that they can also be applied to the DSSE problem. However, execution times for the proposed DSSE method are not reported and the size of the optimization problem has been reduced by dividing the network into several measurement areas.

Multi-area state estimation (MASE) has been studied extensively in the context of transmission systems (Gómez-Expósito et al. 2011) and several applications to DSSE has been proposed in recent years (Džafić et al. 2011; Džafić et al. 2013; Nusrat et al. 2015; Muscas et al. 2015; Muscas et al. 2016; Pau et al. 2017). A distribution network supplied

by one primary substation can contain thousands of MV nodes and tens of thousands of LV nodes. It is clear that these kinds of dimensions put a huge computational burden on the DSSE algorithms. The main purpose of MASE is to speed up the computation by dividing the estimation problem into several sub-problems. This is beneficial in many aspects. In WLS estimation, for example, the execution time of several computationally heavy processes (e.g. gain matrix inversion and Jacobian matrix calculation) depends exponentially on the size of the network. Assuming a constant sub-network size, the time spent on these processes can be linearized with MASE. It also enables parallel computation and distributed state estimation. The downside is that MASE usually leads to some degradation in the estimation accuracy, because all the available measurements are not processed simultaneously (Pau et al. 2017).

5.2 The choice of state estimator

When a DSSE algorithm is coded from scratch, subjected to comprehensive testing, and used in several simulated case studies and real-life smart grid demonstrations, some practical issues should be considered when selecting the underlying DSSE method. Firstly, the selected DSSE method should have a proven record of accomplishments, be easy to implement and understand, and be computationally efficient and robust. These requirements rule out most of the methods presented in Subsection 5.1.2. Many of these methods have been presented only in a few academic papers while the WLS methods presented in Subsection 5.1.1 have been applied in hundreds of different studies. In addition, the PSO-based DSSE methods are too slow for real-time applications and execution time for the interior point optimization based method is unknown. The NN-based methods are robust but not intelligible enough due to their black-box nature. In certain situations, the WLS-methods are known to suffer from convergence issues but this can be alleviated by using robust DSSE as has been done, for example, in (Hayes et al. 2015).

Secondly, the selected DSSE method should be able to handle all measurement types and network configurations typically found in modern smart distribution grids. We have already deduced that a WLS-based DSSE method is the safest bet but we still need to make a choice between node voltage and branch current based DSSE algorithms. They both have their relative strengths. The node voltage method is more established, calculates strongly meshed networks, and handles voltage measurements efficiently. The branch current method has been designed specifically for distribution networks, is faster (Baran & Kelley 1995; Teng 2002; Abdel-Majeed 2016; Primadianto et al. 2016), calculates radial and weakly meshed networks, and handles current measurements efficiently. Both methods are applicable to calculating three-phase MV and LV networks and can handle all types of measurements. Ultimately, the branch current based method was selected due to its faster execution time. The branch currents can be expressed either in rectangular (Baran & Kelley 1995) or in polar form (Wang & Schulz 2004). The polar form was chosen, because then the current magnitude measurements have direct counterparts in the state vector. This not only simplifies current measurement handling but also enables the extraction of branch current magnitude variances directly from the inverted gain matrix.

While most of the other DSSE methods presented in Subsection 5.1.2 were rejected, the MASE approach was determined useful and will be used in this thesis whenever there are distribution networks with clear *natural* measurement areas, such as feeders with current or power flow measurements at the beginning of the feeder. However, feeders are not divided into further measurement areas, even if they have mid-feeder current or power flow measurements. Although a branch current based WLS algorithm was selected in this thesis, it should be noted that with a typical distribution network measurement setup and grid topology, the same state estimation accuracy could also have been achieved with the other WLS-based methods.

5.3 The developed state estimator

The DSSE algorithm development was started from the basic WLS formulation presented in Subsection 5.1.1. Measurement functions and Jacobian matrix entries, which are partial derivatives of the measurements functions with respect to the state variables, were constructed according to the example given in (Wang & Schulz 2004). In [P1], equality constraints were added so that the zero-injection measurements could be forced to zero without using very high measurement weights. The use of equality constraints improved the gain matrix condition number and made the algorithm more robust.

In [P2], bad data detection based on the *largest normalized residual* r_{max}^{N} –test was added and tested both in RTDS simulation environment and in a real-life demonstration. The tests were done on a MV network with a typical measurement configuration, meaning that real-time measurements only from the substation and from the production unit were available and existing load profiles were used as pseudo-measurements for all the other nodes. Later in [P3], the same DSSE algorithm was applied for LV network state estimation and more RTDS tests were conducted. In these tests, it was assumed that all the consumption nodes are monitored in real-time and the measurement reading frequency and averaging time were varied to find out their effect on the estimation accuracy.

Finally, publication [P9] combined the developed AMR-based load profiles and DSSE. State estimation simulations were done on a large distribution network containing both MV and LV networks and estimation accuracy with different types of pseudo-measurements was studied. Furthermore, the developed state estimator was used in simulations and real-life demonstrations done during the INTEGRIS and IDE4L projects.

This section presents the developed algorithm, adds some new properties needed in smart grid environment, shows how DSSE can be integrated into decentralized smart grid monitoring and control concept, and introduces the state forecaster concept. The detailed formulation of measurement equations and Jacobian matrices is omitted from this thesis, instead, the interested readers are referred to (Wang & Schulz 2004) and to IDE4L deliverable (Mutanen et al. 2015). The formulation for equality constrained WLS estimation and bad data detection can be found in [P1] and [P2].

5.3.1 State estimate uncertainties

In smart grid monitoring and control it is often useful to know the uncertainties for the estimated network states. The DSSE accuracy can vary, for example when real-time measurements go off-line due to communication failure or some other malfunction, and the control algorithms must adapt to these changes. Higher uncertainties in the estimated network states mean that safety margins in the network control must be increased.

We have already learned that the state variable variances can be found in the diagonal of the inverted gain matrix. If variances for the other estimated states are needed, some additional work is required. A Jacobian matrix containing partial derivatives for all those states for which we wish to calculate variances must be formed. After the new Jacobian matrix has been formed, the variances can be calculated using (18) (Li 1996). This calculation needs to be done only once after the WLS estimation has converged.

$$var(\boldsymbol{o}(\boldsymbol{y})) = diag(\boldsymbol{K}\boldsymbol{G}^{-1}\boldsymbol{K}^{T}), \tag{18}$$

where $\boldsymbol{o}(\boldsymbol{y})$ is a vector of network state functions
$\boldsymbol{K}$ is the Jacobian of $\boldsymbol{o}(\boldsymbol{y})$
$\boldsymbol{G}$ is the gain matrix used in WLS estimation ($\boldsymbol{H}^{T}\boldsymbol{R}^{-1}\boldsymbol{H}$).

5.3.2 Estimation of weakly meshed networks

Traditionally, distribution networks have been operated radially but meshed operation is expected to increase in future smart grids. The need to connect more DG to the existing networks is driving this development and the advances in protection and distribution automation enable it. In this environment, the state estimator must be able to handle meshed networks. Despite this development, the distribution networks are expected to remain weakly meshed, i.e. the number of meshes remains modest and they are mainly formed by two adjacent feeders operated in a ring.

The branch current based WLS estimators can be modified to calculate weakly meshed networks (Baran & Kelley 1995; Lin et al. 2001; Pau et al. 2013). In a presence of a network loop, nodes can be fed from either direction and additional equations are needed to determine the current flow directions and magnitudes. Kirchhoff's voltage law states that the directed sum of voltages around a closed loop must be zero. This constraint can be added to the WLS formulation either as a virtual measurement or as an equality constraint. The voltage around the loop consist of branch voltage losses, which can be expressed as a product of branch currents and impedances, and Kirchhoff's voltage law can be formulated as:

$$\sum_{j\in\Lambda} \lambda_j \bar{Z}_j \bar{I}_j = 0, \tag{19}$$

where Λ is the set of branches forming the loop, $\bar{Z}_j$ and $\bar{I}_j$ are the impedance and current phasors of the jth branch, and λ_j is +1 or −1 depending on the loop reference direction and on which side of the loop break point the branch is (Pau et al. 2013). The use of λ-parameters is clarified in Figure 5.1. The virtual loop break point needs to be opened temporarily, when the forward sweep method is used to calculate the node voltages, but

the branch currents can be estimated while the loop is closed. The break point location and reference direction can be chosen freely.

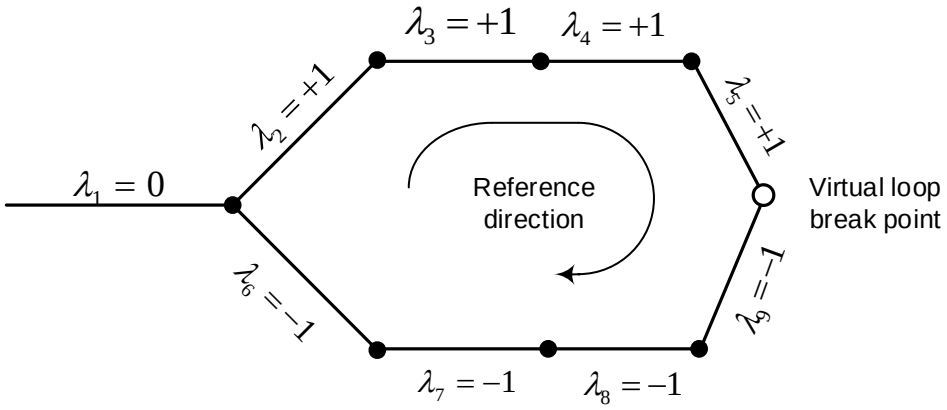


Figure 5.1. Example of a network loop, break point, reference direction and λ values.

Equation (19) is complex and both the real and imaginary parts must be zero. Therefore, this constraint can be divided into two equations corresponding to the real (20) and imaginary (21) parts.

$$c_r = \sum_{j \in \Lambda} \lambda_j |\bar{Z}_j||\bar{I}_j| \cos(\alpha_j + \beta_j) \tag{20}$$

$$c_x = \sum_{j \in \Lambda} \lambda_j |\bar{Z}_j||\bar{I}_j| \sin(\alpha_j + \beta_j) \tag{21}$$

Here α_j and β_j are the current and impedance phasor angles for branch j. The corresponding Jacobian entries—partial derivatives with respect to the state variables—are for (20):

$$\begin{cases} \dfrac{\partial c_r}{\partial |\bar{I}_j|} = \lambda_j |\bar{Z}_j| \cos\left(\alpha_j + \beta_j\right), & \text{if } j \in \Lambda \\ \dfrac{\partial c_r}{\partial |\bar{I}_j|} = 0, & \text{if } j \notin \Lambda \end{cases} \tag{22}$$

$$\begin{cases} \dfrac{\partial c_r}{\partial \alpha_j} = -\lambda_j |\bar{Z}_j||\bar{I}_j| \sin\left(\alpha_j + \beta_j\right), & \text{if } j \in \Lambda \\ \dfrac{\partial c_r}{\partial \alpha_j} = 0, & \text{if } j \notin \Lambda \end{cases} \tag{23}$$

and for (21):

$$\begin{cases} \dfrac{\partial c_x}{\partial |\bar{I}_j|} = \lambda_j |\bar{Z}_j| \sin\left(\alpha_j + \beta_j\right), & \text{if } j \in \Lambda \\ \dfrac{\partial c_x}{\partial |\bar{I}_j|} = 0, & \text{if } j \notin \Lambda \end{cases} \tag{24}$$

$$\begin{cases} \dfrac{\partial c_x}{\partial \alpha_j} = \lambda_j \left|\bar{Z}_j\right| \left|\bar{I}_j\right| \cos\left(\alpha_j + \beta_j\right), & \text{if } j \in \Lambda \\ \dfrac{\partial c_x}{\partial \alpha_j} = 0, & \text{if } j \notin \Lambda \end{cases} \tag{25}$$

5.3.3 Algorithm implementation

The developed state estimator was written into a Matlab program and the computer simulations and real-life demonstrations in [P1]–[P3], and [P9] were performed using Matlab. The tests and real-life demonstrations in the INTEGRIS and IDE4L projects were done using Octave. Octave is an open source Matlab clone and it can often run Matlab code without or with very little modifications. Matlab and Octave are easy to use and provide adequate performance for demonstration purposes (q.v. Section 5.4).

Flow chart of the developed WLS estimator is shown in Figure 5.2. The WLS estimator has been implemented as a Matlab function and it has the following inputs:

- Bus matrix, which contains the bus numbers, initial voltages, load and production measurements or pseudo-measurements and their variances
- Line matrix, which contains start and end nodes, impedances, and capacitive susceptances for each line section
- Power and current flow measurements, current injection measurements, node voltage measurements, and their locations and variances.

The inputs are given in per units and consequently also the outputs are in per units. Most of the outputs are complex numbers, meaning that real and reactive powers can be discriminated and node voltage angles with respect to the slack bus voltage are estimated. The outputs are:

- Node voltage estimates
- Line current flow estimates
- Line power flow estimates
- Line power loss estimates
- Power injection estimates (i.e. load and production estimates)
- Variances for line current flow estimates
- Variances for other selected variables (q.v. Subsection 5.3.1).

The numbered steps in Figure 5.2 have been described below.

1. Input validity check: Rough errors in the input measurements are filtered out using simple logical rules. For example, negative current magnitude measurements and node voltage measurements that are twice as large as the nominal voltage are labelled as bad data and are removed.
2. Branch current calculation: Initial branch currents are calculated using the load and production values provided as inputs. Backward sweep algorithm is used to calculate the branch currents from the bottom up. The lines are modelled with π-model and the algorithm gives separate values for currents at the beginning,

middle, and end of the line. If the network contains loops, virtual loop break points are added so that the network becomes radial and the backward/forward sweep algorithm can be used to calculate the initial branch currents and node voltages.

3. Node voltage calculation: Node voltages are calculated from the top down using the forward sweep method. After this, if virtual loop break points have been added, small initial branch currents flowing from higher to lower potential are added to the lines containing the break points. Zero values on the break lines would later cause a singular gain matrix.
4. Covariance matrix formation: The measurement covariance matrix is formed from the input measurement variances. The measurements are assumed to be uncorrelated. The covariance matrix is a diagonal matrix and the diagonal elements correspond to the accuracy of each measurement (pseudo-measurements included).
5. Measurement vector formation: The provided measurements are collected into a measurement vector.
6. State variable vector formation: The state variable vector is formed from the previously calculated branch current magnitudes and angles. The currents at the beginning of the line are selected as state variables.
7. Jacobian matrix calculation: Jacobian matrices for measurements and equality constraints are calculated. In addition, the measurement function and equality constraint function values are calculated.
8. Calculation of $\Delta \boldsymbol{y}$: Corrections to the state variable vector are calculated by using the Lagrange method presented in [P1].
9. State variable vector update: The state variable vector is updated by adding the previously calculated corrections to it.
10. Mid-line current calculation: Currents in the middle of each line are calculated by adding the appropriate charging currents to the currents presented by the state variables.
11. Node voltage calculation: The node voltages are recalculated using the forward sweep.
12. Bad data detection: Once the largest value in vector $\Delta \boldsymbol{y}$ falls below the pre-set threshold ε, the algorithm exits from the first loop and starts the bad data detection. If bad data is detected, it is removed and the algorithm returns to step 4. If the WLS objective function $J(\boldsymbol{y})$ does not decrease after four iterations, convergence is secured by removing all redundant measurements. The network is fully observable if all load and production nodes have measurements or pseudo-measurements. Since in this thesis all nodes are assumed to have at least pseudo-measurements, all the other measurements can be considered redundant and can be removed temporarily. The calculation returns to step 2 and after the convergence has been achieved the redundant measurements are restored and subjected to bad data detection.
13. Output calculation: Variances for the estimated current magnitudes and other selected variables are calculated using the method described in Subsection 5.3.1. Also, the power flows, injections and losses are calculated.

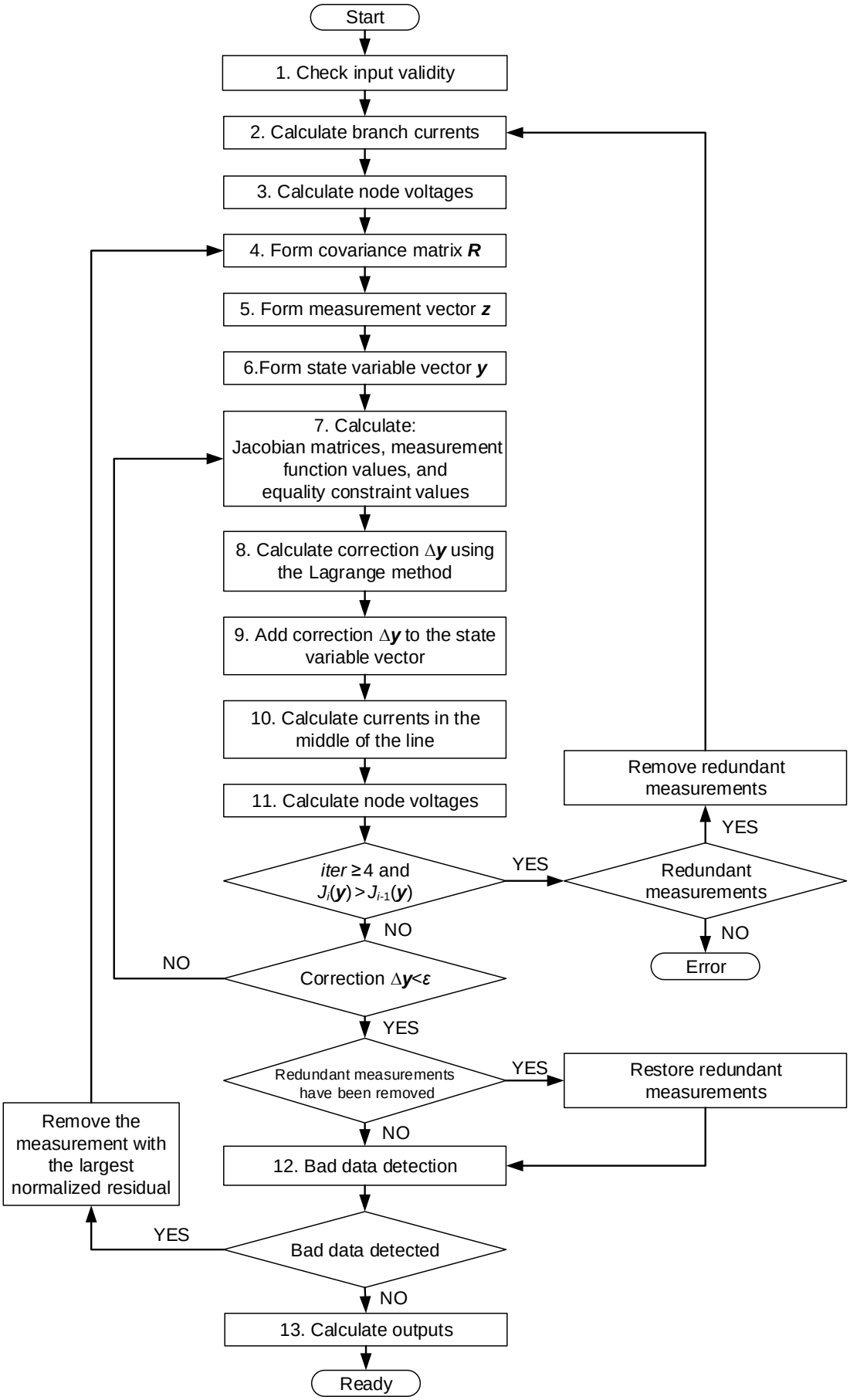


Figure 5.2. Flow chart for the developed WLS estimator.

Only the core of the developed DSSE algorithm is shown in Figure 5.2. In real-life implementations, this state estimator core needs to be supported by several additional functions that do the network topology import, real-time measurement reading, pseudo-measurement reading, etc. In addition, the adequacy of the inputs needs to be checked. For example, we need to check that the received measurements are enough to gain full observability. In IDE4L project, the observability was ensured by using load profiles as back-up pseudo-measurements.

Flow chart of the main script calling the state estimation core and the most important support functions is shown in Figure 5.3. In this simplified example, the state estimator reads all the input information from a database and writes all the state estimation results to the same database. Similar main script structure was used in the IDE4L project to implement both the MV and LV network state estimators.

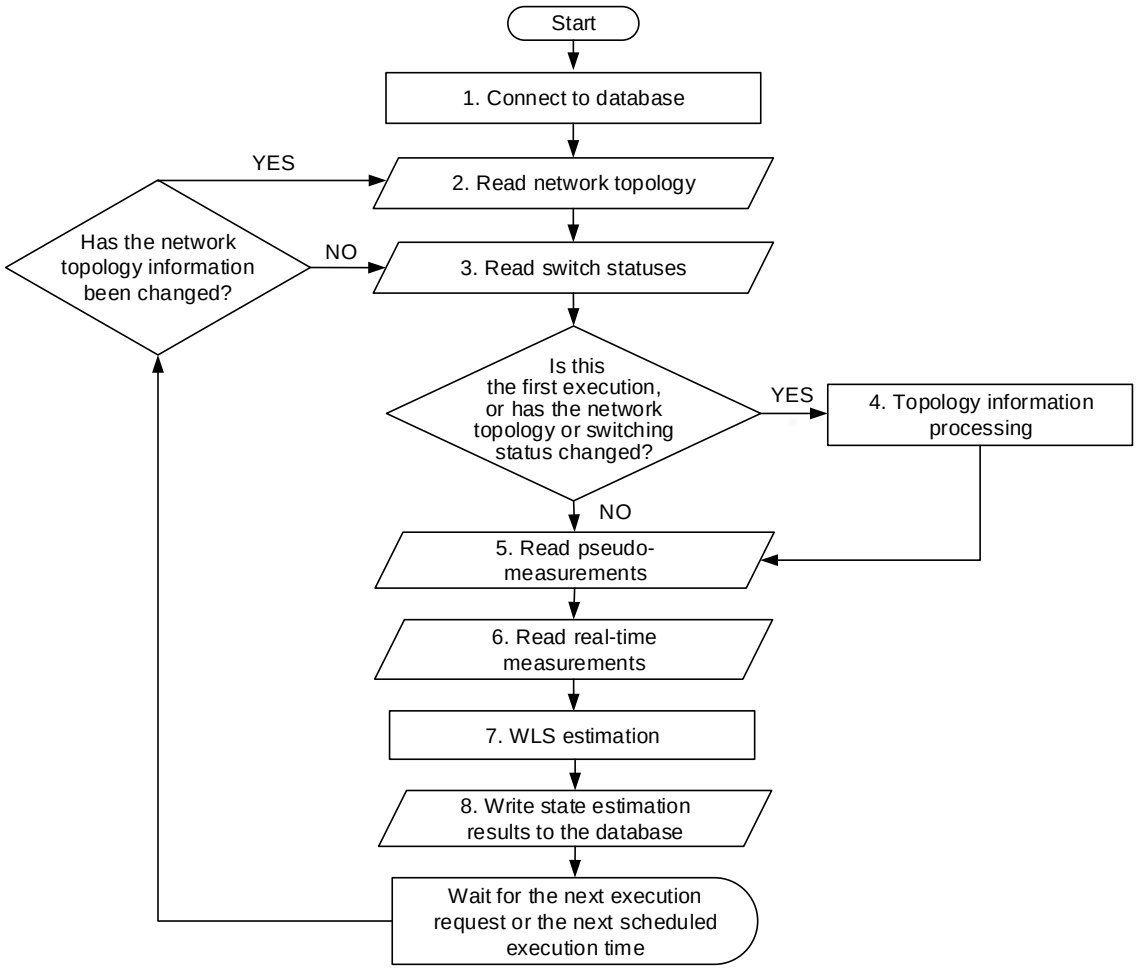


Figure 5.3. Flow chart for the main script calling the WLS estimator.

In [P1], [P2] and [P9], the DSSE simulations were done assuming that all the loads are balanced and the state estimation calculation can be done using an equivalent single phase network model. In [P3] and in the INTEGRIS and IDE4L projects, the networks contained unbalanced loads and three-phase calculation was necessary. The calculation was done assuming that mutual impedances between the phases are zeros and the phases can thus be considered decoupled. It was recognized that this assumption brings some inaccuracy

to the calculation, but since the mutual impedances are much smaller than the self-impedances, this error was assessed to be negligible. In addition, the line mutual impedances, line spacing, and line configuration in the demonstration networks were unknown and therefore an accurate three-phase calculation would not have been possible anyway. In LV networks, the lack of grounding impedance information added even more uncertainty to the calculations. The selection of load model type (constant power, constant impedance or constant current) can also have a significant effect on the power flow calculation results (Ciric et al. 2003). The load model optimization in this regard was outside the scope of this work and simple constant power loads were used throughout the thesis.

5.3.4 Decentralized DSSE

The INTEGRIS and IDE4L projects relied heavily on a decentralized control architecture (Repo et al. 2011), where many distribution network monitoring, control and communication functionalities are distributed to primary and secondary substations. The developed state estimator complies with this concept and can be operated in a decentralized manner. The state estimator benefits from this architecture so that individual networks remain small and they can be estimated quickly and the estimation can be parallelized. Another, perhaps even greater benefit, is that the decentralized system can handle more real-time measurements and, as we know, more real-time measurements lead to more accurate state estimates.

In the decentralized control architecture, real-time measurements are sent to the closest substation automation unit where they are stored, analyzed, and used in local network monitoring and control. The secondary substation automation unit (SSAU) contains the LV network state estimator (LVSE), the load and production forecaster, the state forecaster, and the power controller that do the LV network monitoring and control on a local level. When the LV network monitoring and control are done locally, there is no need to send real-time smart meter measurements to the upper level controller. Only switch and breaker status information, aggregated loading estimates and forecasts, and problem indicators are sent to the higher level system, which in this case is the primary substation automation unit (PSAU). Again, only a fraction of the measured and analyzed data is transferred from PSAU to the next control level, which is at the control centre.

In a decentralized control architecture, the real-time data collection can be based on local communication technologies (e.g. power line communication) and the meter reading frequencies can be high as the number of measurements received by a single substation automation unit is small compared to a situation where all measurements are sent to the control centre. More information on how the substation automation unit design and implementation was done in the IDE4L project is available in (Angioni et al. 2017).

The coordination between LV and MV network state estimators is shown in Figure 5.4. The LV network state estimator (LVSE) outputs phase-wise power flows for the distribution transformer secondary side and a separate transformer model is used to calculate the phase-wise primary side power flows while taking into account the transformer winding configuration and other parameters. All the LVSE results are stored

to the SSAU database where they are available for the network control functions such as the power controller, which implements the coordinated power and voltage control. The SSAU sends the estimated power flows on the primary side of the distribution transformer to the PSAU where they are available for the MV network state estimator (MVSE). In a normal operation mode, state estimate information flows only from SSAU to PSAU but in case of measurement malfunctions, this flow can be reversed. For example, if the secondary substation voltage measurement is missing, the last voltage estimate from MVSE can be sent to SSAU and used in LVSE.

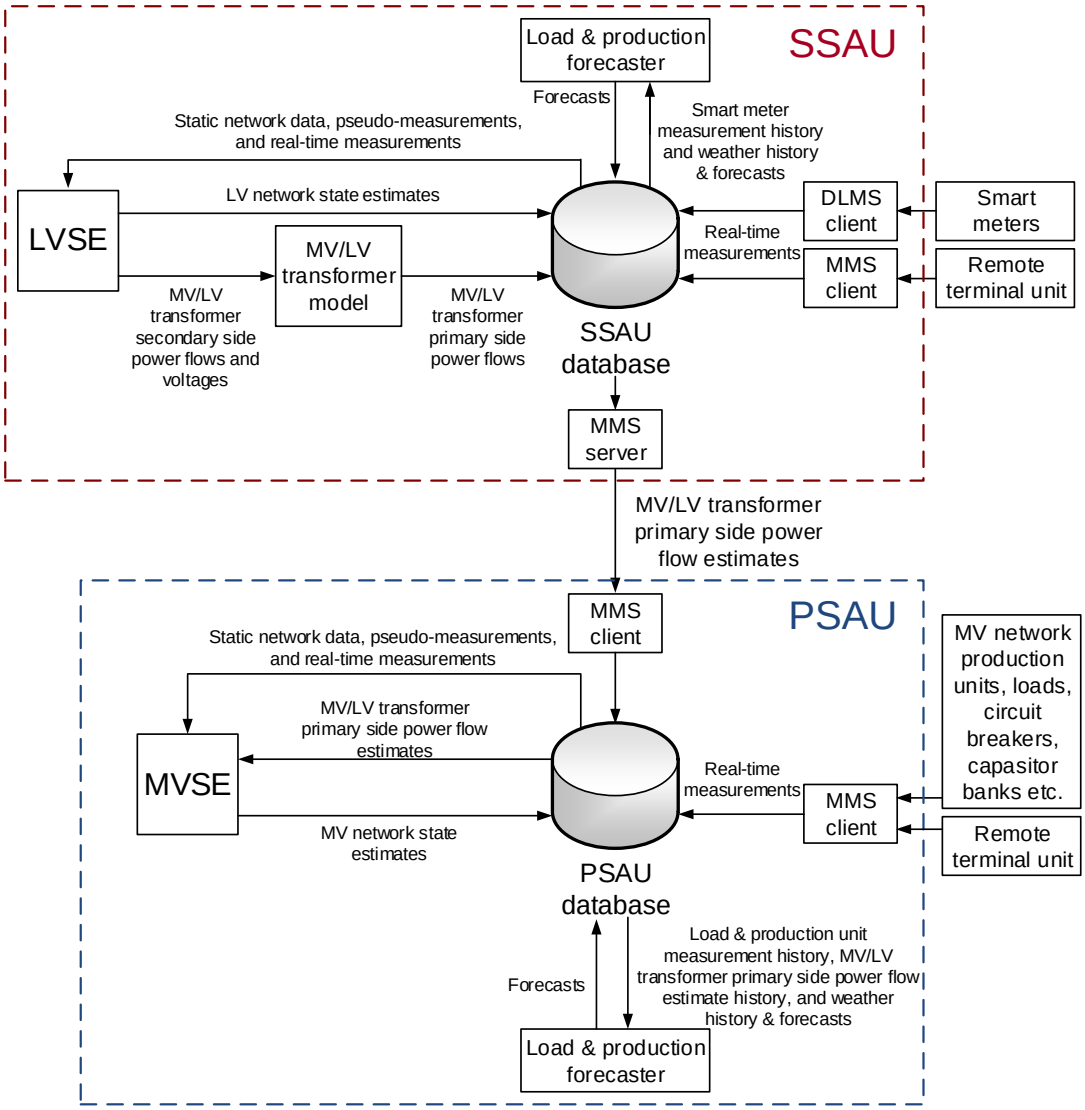


Figure 5.4. Data transfer between LV and MV network state estimators in normal operation mode.

In the IDE4L project, a dedicated load and production forecaster was used to supply the state estimator with load and production pseudo-measurements (Mutanen et al. 2015). The forecaster was made by Daniel Olmeda and Ricardo Vázques in Charles III University of Madrid. Short-term (i.e. 24–48 hours ahead with one hour resolution) load and production forecasts were made using autoregressive models that take into account the measurements history and weather forecasts. Out of these forecasted load and production values, pseudo-measurements for the present moment could always be chosen.

Even if all the load and production points are measured in real-time, the pseudo-measurements are needed as a backup.

The AMR-based load profiles developed in Chapter 4 can also be used with the decentralized DSSE. The clustering and load profile formation are done in a centralized manner but once the load profiles have been uploaded to the substation automation units, they can be used independently. Compared with the load and production forecaster, this approach would have lower computation need as the load profile updating is done less frequently and the behavior of clusters is modelled instead of individual customers. On the down side, the proposed load profiling method would react slower to changes in load and production behavior.

The decentralized control architecture is so light that the substation database and all the monitoring and control functions can be installed into a small industrial PC. Figure 5.5 shows the SSAU used in Unareti S.p.A. demonstration in Italy. Open source software was used to demonstrate that also the software costs can be kept at a minimum. The substation automation units were running Ubuntu Linux operating systems and PostgreSQL databases. The monitoring and control functions were mostly implemented using Octave and Python.

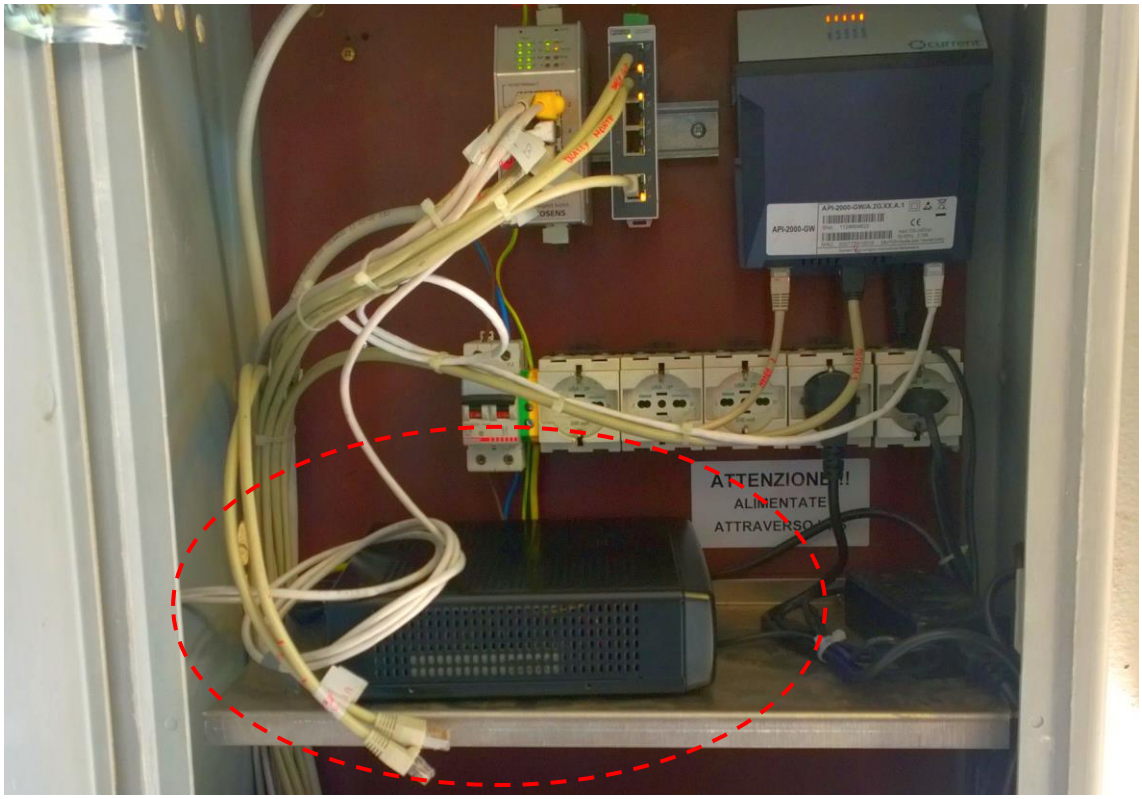


Figure 5.5. SSAU operating in the Unareti S.p.A demonstration network in Brescia, Italy.

5.3.5 State forecasting

In distribution network control, it is often beneficial to know the forthcoming network states. This can help the control system to solve the network congestions and voltage problems even before they happen. In addition, the control functions can better optimize

the use of available resources and unnecessary fluctuations, such as temporary changes in the primary transformer tap position, can be minimized.

In the IDE4L project, a state forecaster was developed based on the above presented state estimator. The main difference between the state forecaster and the state estimator is that the state estimator uses both real-time measurements and forecasted load and productions values for the present time period as inputs, whereas the state forecaster uses only forecasted load and production values for the forthcoming time periods. The state forecaster is run less frequently (e.g. once per hour) but when it is run, it calculates state forecasts for the whole forecasting horizon using the supplied load and production forecast resolution. The load and production forecasts are provided by the load and production forecaster. As in the case of state estimation, AMR-based load profiles could have been used with the state forecasters. In [P8], it is shown that the developed AMR-based load profiles are applicable to short-term load forecasting.

The developed state forecaster uses the same WLS algorithm as the state estimator. The main script calls the WLS algorithm as many times as there are time periods in the load and production forecasts. Between the calls, the main script checks if there are any changes in the scheduled switch statuses and updates the network configuration if needed. In the IDE4L project, only the LV network state forecaster was implemented and the secondary substation voltages were forecasted locally in the SSAU. If also the MV network state forecaster exists, the forecasted secondary substation voltages are sent from PSAU to SSAU, similarly as in the case of state estimation and missing secondary substation voltage measurements.

With the provided inputs, also an ordinary load flow calculation algorithm could have been used to calculate the state forecasts. Although the developed WLS algorithm is slow as a load flow calculation engine, it has some benefits. It can handle redundant (from the observability point of view) forecasts and it can supply variances for the forecasted network states.

5.4 Review and discussion on the achieved results

The basis of the proposed DSSE algorithm was developed in [P1] where Matlab simulations with different real-time measurement configurations were performed. The simulations showed that it is possible to halve the error in MV network state estimation by monitoring only a fraction of the load nodes. In addition, the simulations proved that the state estimation accuracy can be improved significantly by utilizing the voltage measurements. However, the simulations were done assuming a voltage measurement accuracy of ±0.2 %, and sensitivity studies revealed that the benefit of voltage measurements decreases rapidly as the voltage measurement accuracy decreases. This reduces the applicability of voltage measurements especially in LV networks where the less accurate smart meters are expected to supply the voltage measurements.

It was also discovered that the pseudo-measurement accuracies have a big influence on the state estimation accuracy. This finding motivated the author to seek alternative ways to improve the state estimation accuracy. After all, real-time measurements are not cheap.

In addition to meter installations, a communication system capable of transmitting real-time data to the state estimator must be built. If the non-real-time AMR measurements could be used to improve the pseudo-measurement, the state estimation accuracy could be improved without real-time measurement infrastructure.

In [P2], bad data detection was added to the developed state estimator. The revised state estimator was tested with RTDS simulations and in a real-life demonstration. The bad data detection proved to be difficult because there is not enough measurement redundancy in the present distribution networks. With a typical measurement configuration, we can only detect that there is bad data somewhere but we cannot identify if it is in the feeder measurement or in the load (pseudo-)measurements. To solve this issue, we assumed that only the feeder measurements can have bad data and in case of failed convergence, calculated the normalized residuals based on the network states estimated solely based on the pseudo-measurements. In the real-life demonstration, the bad data detection failed because the existing pseudo-measurements could not model the loads accurately in the exceptionally warm weather that occurred during the demonstration. This again motivated the author to develop better load profiles, especially load profiles with better temperature dependency models.

In [P3], the developed LV network state estimator and the needed data collection infrastructure were tested with RTDS simulations. The simulations prepared us for the following real-life demonstrations and showed that the meter reading frequency and length of the measurement averaging period have a big effect on the state estimation accuracy. Later in the IDE4L project, the decentralized DSSE was tested with RTDS simulations and the LV network state estimator was demonstrated in three real-life demonstrations in three different countries. The results of these simulations and demonstrations are reported in (Barbato et al. 2016).

In the IDE4L project, the focus was on the demonstrations. A lot of coordination had to be done to ensure that the input data meets the DSSE needs and the outputs satisfy the needs of the control algorithms. Accuracy-wise, the evaluation of the DSSE performance was difficult in the demonstrations. The actual network states were unknown and there were many possible error sources besides the state estimator, for example, the network model and its parameters, the real-time measurements, and the pseudo-measurements. During laboratory simulations, the execution times for the LV network state estimator were 0.9 and 6.3 seconds for three-phase 15 and 271 bus networks, respectively (Angioni et al. 2017). With the selected open source programs (i.e. Octave and PostgreSQL), writing the results to the SSAU database was the slowest operation and took about 60 % of the total execution time. The actual WLS estimation took only 0.277 and 1.3 seconds for the above-mentioned 15 and 271 bus networks. During the demonstrations, where the networks had 36–271 three-phase buses, the total execution times varied between 20 and 60 seconds, depending on the network size and the SSAU hardware. The SSAU computers used in the demonstrations had low power processors (e.g. Intel Atom) and they were running several functions simultaneously. The laboratory simulations were performed with mid-spec desktop computers.

Matlab and Octave are great environments for algorithm development and prototyping. They are easy to learn, contain extensive libraries of predefined functions and toolboxes, and included tools, such as the integrated editor/debugger, workspace browser, and online documentation, make the programming easy. Matlab and Octave programs do not need to be compiled separately. The possibility to execute code line by line, or in larger chunks, is especially useful when debugging. However, Matlab and Octave are not particularly efficient (excluding matrix operations and liner algebra) and the computation speed could be improved significantly by using lower level programming languages (Andrews 2012), such as Fortran, C or C++.

Publications [P4]-[P8] concentrated on improving the load profiles by using AMR data. The emphasis was on load profiles that could be used in the distribution network calculation: state estimation, operation planning, and short term network planning. The results of this work have been reviewed in Chapter 4.

Finally, in [P9] the developed AMR-based load profiles were combined with DSSE and the estimation accuracy with different types of pseudo-measurements was studied. As expected, the new AMR based load profiles provided better state estimation accuracy than the existing Sener profiles. Depending on the simulation case and estimated variables, the improvements in average estimation accuracy were between 20 and 49 %. Four alternative methods for creating AMR-based load profiles were compared: updated load profiles, cluster profiles, individual profiles, and transformer profiles. In MV network state estimation, the best results were achieved by combining cluster profiles and individual load profiles. In LV network state estimation, the best results were achieved with individual load profiles. However, taking into account the load model complexity and the small differences in accuracy, the combination of cluster profiles and individual profiles might be the best option also for LV networks. Surprisingly good results were achieved when the load allocation was done in relation to the previous year total energy. This approach outperformed both the previous year AMR measurements and the Sener profiles. Thus, an efficient solution was found also for those DSOs that do not have AMR meters.

6 Conclusions and future research

This thesis studies load profiling and distribution system state estimation. At first glance these two topics may appear independent, but if one takes a closer look, the similarities and interconnections are easy to see. In both cases, the goal is to produce accurate load estimates by leveraging the available input data. In case of load profiling, the input data contains historical interval measurements and in the case of DSSE, the input data contains a network model, real-time measurements and pseudo-measurements. The difference is that the DSSE estimates also other variables, for example, line power and current flows and node voltages, which happen to depend on the network loading. The accuracy of the DSSE depends mainly on the quality, quantity, and location of the real-time measurements and on the quality of the pseudo-measurements. So, one way to enhance the DSSE accuracy is to improve the pseudo-measurements. And what are the pseudo-measurements? They are initial load estimates based on load profiles or some other similar load models. Thereby, improving the load profiles, also improves the DSSE accuracy.

This thesis uses AMR measurements (i.e. hourly or half-hourly interval measurements) to improve the load profiles. Early on it was observed that the load profiling accuracy can be improved by clustering customers into similarly behaving groups and by creating new cluster specific load profiles. Updating the existing customer class load profiles did not provide as good results as the cluster profiles, mainly because the existing customer classification was inaccurate and could not be trusted.

A two-stage clustering method that includes temperature dependency calculation, normalization, dimension reduction, cluster weighting, outlier filtering, and selection of customers for individual load profiling was developed and many different clustering algorithms were tested. Good news for anyone following in my footsteps is that, in this application, the best results were achieved with a simple and well-known *k*-means algorithm. Many of the more advanced clustering algorithms failed to achieve similar accuracy or clustering speed. The bad news is that the number of clusters must be known a priori in *k*-means and, based on the accuracy index comparisons made in this thesis, it is not possible to determine the optimum number of clusters unambiguously.

Most of the clustering studies in this thesis are done with an assumption that the customers are modelled with yearly load profiles that are compatible with the existing Finnish load profile format. This enables easy and fast implementation as the existing distribution system software can utilize the developed AMR-based load profiles with little or no changes. In many other countries, typical daily profiles are used instead of yearly load profiles. The load profiling procedure presented in Chapter 4 is not directly applicable in these countries, but several parts of the developed two-stage clustering method could also be used with typical daily profiles. Additional methods for constructing customer level load models from the resulting cluster models would be needed though. Some methods

for clustering daily profiles and for composing customer level load models are presented in [P7], but further research is undoubtedly needed.

A distribution system state estimator was built during this thesis work and while many aspects of state estimation were studied, the most important goal was to prove that the new AMR-based load profiles improve the accuracy of DSSE. This goal was achieved by conducting comprehensive simulations with real AMR data and 10,433 bus distribution network. Depending on the studied case and used accuracy index, the developed AMR-based load profiles improved the DSSE accuracy by 20–49 % compared with a situation where the presently available customer class load profiles were used.

Given how significant efforts have been made in the literature to develop more accurate DSSE methods and to determine how many and what type of real-time measurements should be installed in what part of the network, it is amazing how little attention pseudo-measurement development has received. Since the DSSE accuracy can be improved this much just by exploiting the existing AMR measurements, the use of AMR based load profiles should be at the top-end of a tool list in all DSOs that wish to improve the DSSE accuracy. Only the deployment of a state estimator and installation of real-time substation measurements should be higher in the priority list. If the desired estimation accuracy is not achieved with these and AMR-based load profiles, then the addition of other real-time measurements could be considered. The AMR measurements are basically *free,* since in many countries they are already collected for billing and other purposes, but the additional real-time measurement devices and the communication infrastructure they require are costly. The main cost component in AMR-based load profiling is the data analysis, which can be automated for the most part.

Some argue that load profiles are obsolete and can be directly substituted with smart meter data. This is not true at all. The studies in this thesis have shown that historical data as such is not suited for forecasting. Data-based load models that take weather forecasts and other external variables into account are needed for accurate load forecasting. Presently, most of the AMR systems collect the interval data only once a day. This is not enough for real-time monitoring and load profiles are also needed in state estimation.

The next big step in smart grid evolution might be the implementation of a decentralized control architecture. This new architecture would enable the real-time smart meter reading and use of local load and production forecasters. These forecasters might be more accurate than the AMR-based load profiles presented in this thesis, but at the expense of higher computational complexity. Even though the technological feasibility of the decentralized control architecture has already been demonstrated, it is to be seen when it becomes economically viable. Having participated in such a demonstration project, the author predicts that this will take many years and the load profiles are needed long in the foreseeable future.

I hope this thesis convinces all the readers that it is beneficial to use clustering in load profiling and the AMR-based load profiles can significantly improve the DSSE accuracy.

6.1 Future research topics

Much was done during this thesis work, but as always in research, new research topics emerged faster than the old ones could be completed. In addition to the time constraints, the lack of necessary measurement data impeded some research ideas. In the future, the developed load profiling and state estimation methods could be improved if the following topics were to be studied:

- In the proposed load profiling method, each cluster can contain different types of customers that just happen to behave similarly. Moreover, in some clustering methods, the cluster numbers can vary randomly. From the usability point of view, it would therefore be good if descriptive names could be (automatically) defined for the clusters.
- The change detection methods developed in (Chen 2014; Nurmiranta 2017) should be further developed and integrated with the proposed load profiling method.
- More accurate load models for certain loads, such as storage heating, should be developed.
- The effect of demand response on different types of customers should be studied and load response models should be developed.
- The accuracy and applicability of the daily load profiles used in [P7] should be compared with the yearly load profiles used elsewhere in this thesis. Defining their relative strengths and weaknesses could help to improve both model types.
- When measurements become available, reactive power profiles and phase-wise load profiles should be studied.
- The accuracy of the proposed load profiling method should be compared with the load and production forecaster developed in the IDE4L project. More comparisons with other state-of-the-art forecasting methods found in the literature should also be made. These comparisons should be made on an individual customer level or with small aggregated customer groups. Large customer groups have already been studied in [P8].
- A truly three-phase DSSE algorithm that takes into account the line mutual impedances should be developed.
- Voltage measurement based phase detection methods should be developed so that the phase-wise measurements and load models can be placed on the correct phase in DSSE. Furthermore, possibilities to use AMR measurements in line parameter estimation and even in network topology estimation could be studied.

References

Abdel-Majeed, A. and Braun, M. (2012) 'Low voltage system state estimation using smart meters', *47th International Universities Power Engineering Conference (UPEC)*, London, U.K., pp. 1–6.

Abdel-Majeed, A., Tenbohlen, S., Schöllhorn, D. and Braun M. (2013) 'Meter placement for low voltage system state estimation with distributed generation', *22nd International Conference and Exhibition on Electricity Distribution (CIRED)*, Stockholm, Sweden, pp. 1–4.

Abdel-Majeed, A. (2016) '*Three-phase state estimation for low-voltage grid*', Ph.D. thesis, University of Stuttgart, Germany.

Abur, A. and Expósito, A.G. (2004) '*Power System State Estimation: Theory and Implementation*', New York, U.S., Marcel Dekker Inc.

Adato (2013) '*Kotitalouksien sähkönkäyttö 2011 (Domestic electricity use 2011)*', Adato Energia Oy, Report number: 26.2.2013. (in Finnish)

Akaike, H. (1974) 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control*, 19(6), pp. 716–723.

Alimardani, A., Therrien, F., Atanackovic, D., Jatskevich, J. and Vaahedi, E. (2015) 'Distribution system state estimation based on nonsynchronized smart meters', *IEEE Transactions on Smart Grid*, 6(6), pp. 2919–2928.

Allera, S.V., Alcock, N.D. and Cook, A.A. (1990) 'Load research in a privatized electricity supply industry', *6th International Conference on Metering Apparatus and Tariffs for Electricity Supply*, Manchester, U.K., pp. 1–5.

Andrews, T., (2012) '*Computation time comparison between Matlab and C++ using launch windows'*, [Online], Available at: http://digitalcommons.calpoly.edu/aerosp/78/, (Accessed 22.12.2017).

Angioni, A., Kulmala, A., Della Giustina, D, *et al.* (2017) 'Design and implementation of a substation automation unit', *IEEE Transactions on Power Delivery*, 32(2), pp. 1133–1142.

Ankerst, M., Breunig, M.M., Kriegel, H-P. and Sander, J. (1999) 'OPTICS: ordering points to identify the clustering structure', *ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA, U.S., pp. 49–60.

Argonne National Laboratory (1980) '*Load research manual – volume 2: fundamentals of implementing load research procedures*', Report number: ANL/SPG-13.

Ari, C., Aksoy, S. and Arikan, O. (2012) 'Maximum likelihood estimation of Gaussian mixture models using stochastic search', *Pattern Recognition*, 45(7), pp. 2804–2816.

Arthur, D. and Vassilvitskii, S. (2007) '*K*-means++: the advantages of careful seeding', *18th ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, U.S., pp. 1027–1035.

Arvidsson, M. (2015) '*Analys av mellanspänningsnätet i centrala delar av Västerås stad (Analysis of medium voltage network in Västerås city centre)*', M.Sc. thesis, Faculty of science and technology, Uppsala University, Sweden. (In Swedish)

Baarsch, J. and Celebi, E. (2012) 'Investigation of internal validity measures for K-means clustering', *20th International MultiConference of Engineers and Computer Scientists (IMECS)*, Hong Kong, China, pp. 1–6.

Bahmani, B., Moseley, B., Vattani, A., Kumar, R. and Vassilvitskii, S. (2002) 'Scalable k-means++', Proceedings of the VLDB Endowment, 5(7), pp. 622–633.

Ball G.H. and Hall D.J. (1965) '*ISODATA, a novel method of data analysis and pattern classification*', Stanford Research Institute, Report number: AD699616.

Baran, M.E. and Kelley, A.W. (1994) 'State estimation for real-time monitoring of distribution systems', *IEEE Transactions on Power Systems*, 9(3), pp. 1601–1609.

Baran, M.E. and Kelley, A.W. (1995) 'A branch-current-based state estimation method for distribution systems', *IEEE Transactions on Power Systems*, 10(1), pp. 483–491.

Baran, M.E., Zhu, J. and Kelley, A.W. (1996) 'Meter placement for real-time monitoring of distribution feeders', *IEEE Transactions on Power Systems*, 11(1), pp. 332–337.

Baran, M.E. and McDermott, T.E. (2009) 'Distribution system state estimation using AMI data', *IEEE/PES Power Systems Conference and Exposition (PSCE)*, Seattle, WA, U.S., pp. 1–3.

Barbato, A., Mutanen, A., Alvarez, A., *et al.* (2016) *'Deliverable 7.2 – Overall final demonstration report*', Ideal Grid for All (IDE4L), Project report.

Barbeiro, P.N.P., Teixeira, H., Krstulovic, J., Pereira, J. and Soares, F.J. (2015) 'Exploiting autoencoders for three-phase state estimation in unbalanced distribution grids', *Electric Power Systems Research*, 123, pp. 108–118.

Batrinu, F., Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M. and Toader, C. (2005) 'Efficient iterative refinement clustering for electricity customer classification', *6th IEEE PES PowerTech Conference*, St. Petersburg, Russia, pp. 1–7.

Bernieri, A., Betta, G., Liguori, C. and Losi, A. (1996) 'Neural networks and pseudo-measurements for real-time monitoring of distribution systems', *IEEE Transactions on Instrumentation and Measurement*, 45(2), pp. 645–650.

Bizzozero, F., Gruosso, G. and Vezzini, N. (2016) 'A time-of-use-based residential electricity demand model for smart grid applications', *16th International Conference on Environment and Electrical Engineering (EEEIC)*, Florence, Italy, pp. 1–6.

Bock, H-H. (2007) 'Clustering methods: a history of *k*-means algorithms'. In: Brito, P., Bertrand, P., Cucumel, G. and Carvalho, F. 'Selected contributions in data analysis and classification', Berlin, Germany, Springer, pp. 161–172.

Branke, J., Dep, K., Dierolf, H. and Osswald M. (2004) 'Finding Knees in Multi-objective Optimization', *8th International Conference on Parallel Problem Solving from Nature (PPSN VIII)*, Birmingham, U.K., pp. 722–731.

Brinkmann, B. and Negnevisky, M. (2016) 'Robust state estimation in distribution networks', *27th Australasian Universities Power Engineering Conference (AUPEC)*, Brisbane, Australia, pp. 1–5.

Carpaneto, E., Chicco, G., Napoli, R. and Scutariu, M. (2006) 'Electricity customer classification using frequency–domain load pattern data', *International Journal of Electrical Power & Energy Systems*, 28(1), pp. 13–20.

CENELEC (2016) '*Smart grids*', [Online], Available at: www.cenelec.eu/aboutcenelec/whatwedo/technologysectors/smartgrids.html, (Accessed 31.10.2016).

Chen, Q., Kaleshi, D., Armour, S. and Fan, Z. (2014) 'Reconsidering the smart metering data collection frequency for distribution state estimation', *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Venice, Italy, pp. 517–522.

Chen, T. (2014) '*Customer behaviour change detection based on AMR measurements*', M.Sc. thesis, Tampere University of Technology, Finland.

Chen, T., Mutanen, A., Järventausta, P. and Koivisto, H. (2015) 'Change detection of electric customer behavior based on AMR measurements', *11th IEEE PES PowerTech Conference*, Eindhoven, Netherlands, pp. 1–6.

Cheng, Y. and Li, Y. (2009) 'Research of classification of electricity consumers based on principal component analysis', *6th International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, China, pp. 201–206.

Chicco, G., Napoli, R., Postolache, P., Scutariu, M. and Toader, C. (2003) 'Customer characterization options for improving the tariff offer', *IEEE Transactions on Power Systems*, 18(1), pp 381–387.

Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M. and Toader, C. (2005) 'Emergent electricity customer classification', *IEE Proceedings - Generation, Transmission and Distribution*, 152(2), pp. 164–172.

Chicco, G. and Ilie, I.S. (2009) 'Support vector clustering of electrical load pattern data', *IEEE Transactions on Power Systems*, 24(3), pp. 1619–1628.

Chicco, G. (2012) 'Overview and performance assessment of the clustering methods for electrical load pattern grouping', *Energy*, 42(1), pp. 68–80.

Chicco, G., Ionel, O-M. and Porumb, R. (2013) 'Formation of load pattern clusters exploiting ant colony clustering principles', *15th IEEE EuroCon - International Conference on Computers as a Tool*, Zagreb, Croatia, pp. 1460–1467.

Ciric, R.M., Feltrin, A.P. and Ochoa, L.F. (2003) 'Power flow in four-wire distribution networks-general approach', *IEEE Transactions on Power Systems*, 18(4) pp. 1283–1290.

Cobelo I., Shafiu A., Jenkins N. and Strbac G. (2007) 'State estimation of networks with distributed generation', *European Transactions on Electrical Power*, 17(1), pp. 21–36.

Dahlström, C., Eriksson, E., Fritz, P. and Lydén, P. (2011) '*Framtagande av effektprofiler samt uppbyggnad av databas över elanvändningen vid kall väderlek (Formation of load profiles and development of database on electricity use in cold weather)*', Elforsk, Report number: 11:12. (in Swedish)

Damavandi, M.G., Krishnamurthy, V. and Martí, J.R. (2015) 'Robust meter placement for State estimation in active distribution systems', *IEEE Transactions on Smart Grid*, 6(4), pp. 1972–1982.

Dang-Ha, T-H., Olsson, R. and Wang, H. (2016) 'Clustering methods for electricity consumers: an empirical study in Hvaler – Norway', *Norsk Informatikkonferanse (NIK)*, Bergen, Norway, pp. 1–12.

Das, I. (1999) 'On characterizing the "knee" of the Pareto curve based on Normal-Boundary Intersection', *Structural optimization*, 18(2–3), pp. 107–115.

Davies, D.L. and Bouldin, D.W. (1979) ‘A Cluster Separation Measure’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), pp. 224–227.

Dep, K. and Gupta, S. (2011) ‘Understanding knee points in bicriteria problems and their implications as preferred solution principles’, *Engineering Optimization*, 43(11), pp. 1175–1204.

Di Nuovo, A.G. and Catania, V. (2008) ‘An evolutionary fuzzy c-means approach for clustering of bio-informatics databases’, *16th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Hong Kong, China, pp. 2077–2082.

Duda, R.O., Hart, P.E., and Stork, D.G. (2012) ‘*Pattern classification*’, 2nd edition, Somerset, NJ, U.S., John Wiley & Sons Inc.

Dunn, J.C. (1973) ‘A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters’, *Cybernetics and Systems*, 3(3), pp. 32–57.

Dzafic, I., Neisius, H.-T. and Henselmeyer, S. (2011) ‘Real time distribution system state estimation based on interior point method’, *17th Power Systems Computation Conference (PSCC)*, Stockholm, Sweden, pp. 1–9.

Džafić, I., Gilles, M., Jabr, R.A., Pal, B.C. and Henselmeyer, S. (2013) ‘Real time estimation of loads in radial and unsymmetrical three-phase distribution networks’, *IEEE Transactions on Power Systems*’, 28(4), pp. 4839–4848.

EC JRC (2016) ‘*Smart metering deployment in the European Union*’, [Online], European Commission Joint Research Centre (EC JRC), Available at: http://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union, (Accessed 14.11.2016).

ELEXON (2013) ‘*Load profiles and their use in electricity settlement*’, [Online]. Available at: www.elexon.co.uk/wp-content/uploads/2013/11/load_profiles_v2.0_cgi.pdf, (Accessed 20.7.2016).

Finnish Energy (2016), ‘*Tuntimittauksen periaatteita (Principles of hourly metering)*’, [Online], Available at: http://energia.fi/files/1153/Tuntimittaussuositus_paiv_2016 1012.pdf, (Accessed 21.12.2016). (in Finnish)

Energy Authority (2017), ‘*Sähköverkon haltijat (Distribution network owners)*’, [Online], Available at: www.energiavirasto.fi/sahkoverkon-haltijat, (Accessed 23.1.2017). (in Finnish)

Engblom, O. and Ueda, M. (2008) ‘*Representativa testnät för Svenska eldistributionsnät (Representative test networks for Sweden)*’. Elforsk, Report number: 08:42. (in Swedish)

ERGEG (2009) ‘*Position paper on smart grids - An ERGEG public consultation paper*’, European Regulators Group for Electricity & Gas, Report number: E09-EQS-30-04.

Ester, M., Kriegel, H-P., Sander, J. and Xu, X. (1996) ’A density-based algorithm for discovering clusters in large spatial databases with noise’ *2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, Portland, OR, U.S., pp. 226–231.

Estivill-Castro, V. (2002) ‘Why so many clustering algorithms – a position paper’, *SIGKDD Explorations*, 4(1), pp. 65–75.

European Council conclusion SN 79/14 of 23–24 October 2014 on 2030 climate and energy policy framework.

European Council directive 2009/28/EC of 23 April 2009 on the promotion of the use of energy from renewable sources.

European Parliament and Council decision 406/2009/EC of 23 April 2009 on the effort of member states to reduce their greenhouse gas emissions to meet the community's greenhouse gas emission reduction commitments up to 2020.

European Parliament and Council directive 2012/27/EU of 25 October 2012 on energy efficiency.

Ferdowsi, M., Löwen, A., McKeever, *et al.* (2014) 'New monitoring approach for distribution systems', *31st IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Montevideo, Uruguay, pp. 1506–1511.

Figueiredo, V., Rodrigues, F., Vale, Z. and Gouveia J.B. (2005) 'An electric energy consumer characterization framework based on data mining techniques', *IEEE Transactions on Power Systems*, 20(2), pp. 596–602.

Fingrid (2017a) '*Electricity retail market business processes in Datahub*', [Online], Available at: www.ediel.fi/sites/default/files/Electricity retail market business processes in Datahub version 1.2.pdf, (Accessed 20.12.2017).

Fingrid (2017b) '*Load and generation*', [Online], Available at: www.fingrid.fi/en/electricity-market/load-and-generation/, (Accessed 13.1.2017).

Flath, C., Nicolay, D., Conte, T., van Dinther, C. and Filipova-Neumann, L. (2012) 'Cluster analysis of smart metering data - An implementation in practice', *Business & Information Systems Engineering*, 4(1), pp. 31–39.

Gerbec, C., Gašperič, S., Šmon, I. and Gubina, F. (2003a) 'Consumers' load profile determination based on different classification methods', *IEEE Power Engineering Society General Meeting*, Toronto, Canada, pp. 990–995.

Gerbec, C., Gašperič, S. and Gubina, F. (2003b) 'Determination and allocation of typical load profiles to the eligible consumers', *5th IEEE PES PowerTech Conference*, Bologna, Italy, pp. 1–5.

Ghahramani, Z. and Hinton, G.E. (1997) '*The EM algorithm for mixtures of factor analyzers*', University of Toronto, Report number: CRG-TR-96-1.

Ghosh, A., Lubkeman, D., Downey, M. and Jones, R. (1997) 'Distribution circuit state estimation using a probabilistic approach', *IEEE Transactions on Power Systems*, 12(1), pp. 45–51.

Gómez-Expósito, A., de la Villa Jaén, A., Gómez-Quiles, C, Rousseaux, P. and Van Cutsem, T. (2011) 'A taxonomy of multi-area state estimation methods', *Electric Power Systems Research*, 81, pp. 1060–1069.

Gondzio, J. (2012) 'Interior point methods 25 years later', *European Journal of Operational Research*, 218(3), pp. 587–601.

Grip, K. (2013) '*Pienasiakkaan kysynnän jouston ja oman tuotannon vaikutukset kuormitusmalleihin (Effect of demand response and own production of small-scale customer on load profiling)*', M.Sc. thesis, Tampere University of Technology, Finland. (in Finnish)

Grip, K., Mutanen, A. and Järventausta, P. (2014) 'Effects of demand response on load profiling of small-scale customers', *11th Nordic Electricity Distribution and System Management Conference (NORDAC)*, Stockholm, Sweden, pp. 1–8.

Guha, S., Rastogi, R. and Shim, K. (2001) ‘CURE: an efficient clustering algorithm for large databases’, *Information Systems*, 26(1), pp. 35–58.

Haben, S., Singleton, C. and Grindrod, P. (2016) ‘Analysis and clustering of residential customers energy behavioral demand using smart meter data’, *IEEE Transactions on Smart Grid*, 7(1), pp. 136–144.

Han, J., Kamber, M. and Pei, J. (2012) ’*Data mining: concepts and techniques*’, 3rd edition, Waltham, MA, U.S., Morgan Kaufmann.

Handschin, E., Langer, M. and Kliokys, E. (1995) ‘An interior point method for state estimation with current magnitude measurements and inequality constraints’, *IEEE Power Industry Computer Application Conference*, Salt Lake City, UT, U.S., pp. 385–391.

Hayes, B.P., Gruber, J.K. and Prodanovic, M. (2015) ‘A closed-loop state estimation tool for MV network monitoring and operation’, *IEEE Transactions on Smart Grid*, 6(4), pp. 2116–2125.

Hemmingson, M. and Lexholm, M. (2013) ‘*Dimensioning of smart power grids for the future*’, Elforsk, Report number: 13:98.

Hinneburg, A. and Keim, D.A. (1998) ‘An efficient approach to clustering in large multimedia databases with noise’, *4th International Conference on Knowledge Discovery and Data Mining (KDD)*’, New York, NY, U.S., pp. 58–65.

Ilango, M.R. and Mohan, V. (2010) ‘A survey on grid based clustering algorithms’, *International Journal of Engineering Science and Technology*, 2(8), pp. 3442–3446.

Jain, A.K. (2010) ‘Data clustering: 50 years beyond K-means’, *Pattern Recognition Letters*, 31, pp. 651–666.

Jalonen, M., Ruska, M. and Lehtonen, M. (2003) *’Kuormitustutkimus 2003 (Load research study 2003)*’. VTT Technical Research Centre of Finland, Report number: PRO1/P7028/03. (in Finnish)

Jia, Z., Chen, J. and Liao, Y. (2013) ‘State estimation in distribution system considering effects of AMI data’, *Proceedings of IEEE Southeastcon*, Jacksonville, FL, U.S., pp. 1–6.

Järventausta, P., Verho, P., Partanen, J. and Kronman, D. (2011) ‘Finnish smart grids – A migration from version one to the next generation’, *21st International Conference on Electricity Distribution (CIRED)*, Frankfurt, Germany, pp. 1–4.

Karali, N., Marnay, C., Yan, T., et al. (2015) ‘Towards uniform benefit-cost analysis for smart grid projects: an example using the smart grid computation tool’, Lawrence Berkeley National Laboratory, Report number: LBNL-1003908.

Karypis, G., Han, E-H. and Kumar, V. (1999) ‘Chameleon: hierarchical clustering using dynamic modeling’, *IEEE Computer*, 32(8), pp. 68–75.

Kim, Y-I., Ko, J-M. and Choi S-H. (2011) ‘Methods for generating TLPs (typical load profiles) for smart grid-based energy programs’, *IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*, Paris, France, pp. 1–6.

Kirjavainen, M. and Seppälä, A. (2007) '*Sähkön pienkuluttajien etäluettavan mittaroinnin päivitetty tila (The updated status of electricity consumption remote reading)*', Enease Oy, Report number: 10/464/2007. (in Finnish)

Koivisto, M., Heine, P., Mellin, I. and Lehtonen, M. (2013) 'Clustering of connection points and load modeling in distribution systems', *IEEE Transactions on Power Systems*, 28(2), pp. 1255–1265.

Koponen, P. and Niska, H. (2016) 'Hybrid model for short-term forecasting of loads and control responses', *6th IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Ljubljana, Slovenia, pp. 1–6.

Koponen P., Saco L.D., Orchard N., *et al.* (2008) '*Definition of smart metering and applications and identification of benefits*', European Smart Meter Alliance (ESMA), Project report.

Kopsakangas-Savolainen, M. (2002) 'Tutkimus sähkömarkkinoiden vapauttamisesta Suomessa (A study on the deregulation of the Finnish electricity markets)', *Kansantaloudellinen aikakausikirja*, 98(1), pp. 74–79. (in Finnish)

Kosonen, A. (2008) '*Power line communication in motor cables of variable-speed electric drives – analysis and implementation*', Ph.D. thesis, Lappeenranta University of Technology, Finland.

Kulmala A, (2014) '*Active voltage control in distribution networks including distributed generation*', Ph.D. thesis, Tampere University of Technology, Finland.

Kulmala, A., Mutanen, A., Koto, A., Repo, S. and Järventausta, P. (2010) 'RTDS verification of a coordinated voltage control implementation for distribution networks with distributed generation', *1st IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Gothenburg, Sweden, pp. 1–8.

Kulmala, A., Mutanen, A., Koto, A., Repo, S. and Järventausta P. (2012) 'Demonstrating coordinated voltage control in a real distribution network', *3rd IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Berlin, Germany, pp. 1–8.

Kryszczuk, K. and Hurley, P. (2010) 'Estimation of the number of clusters using multiple cluster validity indices', *9th International Workshop on Multiple Classifier Systems*, Cairo, Egypt, pp. 1–10.

Lakervi, E. and Holmes, E.J. (2003) '*Electricity distribution network design*', 2nd edition, London, U.K., Peregrinus Ltd.

Larsson, L., Lindskoug, S. and Lindén, M. (2006) '*Elförbrukningens karaktär vid kall väderlek (Electricity consumption characteristics in cold weather)*', Elforsk, Report number: 06:62. (in Swedish)

Li, K. (1996) 'State estimation for power distribution systems and measurement impacts', *IEEE Transaction on Power Systems*, 11(2), pp. 911–916.

Li, R., Li, F. and Smith, N.D. (2016) 'Multi-resolution load profile clustering for smart metering data', *IEEE Transactions on Power Systems*, 31(6), pp. 4473–4482.

Liao, T.W. (2005) 'Clustering of time series data - a survey', *Pattern Recognition*, 38, pp. 1867–1874.

Lin, W.-M. and Teng, J.-H. (1996) ‘State estimation for distribution systems with zero-injection constraints’, *IEEE Transactions on Power Systems*, 11(1), pp. 518–524.

Lin, W.-M., Teng, J.-H. and Chen, S.-J. (2001) ‘A highly efficient algorithm in treating current measurements for branch-current-based distribution state estimation’, *IEEE Transactions on Power Delivery*, 16(3), pp. 433–439.

Lloyd, S. (1982) ‘Least squares quantization in PCM’, *IEEE Transactions on Information Theory*, 28(2), pp. 129-137. Originally an unpublished Bell Laboratories technical note (1957).

Lo, K.L., Zakaria, Z. and Sohod, M.H. (2005) ’Determination of consumers’ load profiles based on two-stage fuzzy c-means’, *5th International Conference on Power Systems and Electromagnetic Compatibility*, Corfu, Greece, pp. 212–217.

Lowry, R. (2017), ‘*Concepts & applications of inferential statistics*’, [Online], Available at: http://vassarstats.net/textbook/, (Accessed 17.1.2017).

Lu, C.N., Teng, J.-H. and Liu, W.-H. (1995) ‘Distribution System State Estimation’, *IEEE Transactions on Power Systems*, 10(1), pp. 229–240.

Lu, Y., Zhang, T. and Zeng, Z. (2016) ‘Adaptive weighted fuzzy clustering algorithm for load profiling of smart grid customers’, *5th International Conference on Communications in China (ICCC)*, Chengdu, China, pp. 1–6.

Luke (2016) ‘*Statistics database*’, [Online], Natural Resources Institute Finland (Luke), Available at: http://statdb.luke.fi/PXWeb/pxweb/en/LUKE/, (Accessed 28.12.2016).

MacQueen, J. (1967) ‘Some methods for classification and analysis of multivariate observations’, *5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, U.S., pp. 281–297.

MarketMath (2016), ‘*Own profiles*’, [Online], Available at: www.marketmath.info/?q=node/17, (Accessed 9.9.2016).

MathWorks (2017) ‘*Manova1*’, [Online], Available at: https://se.mathworks.com/help/stats/manova1.html, (Accessed 20.12.2017).

McLachlan, G.J., Peel, D. and Bean, R.W. (2003) ‘Modelling high-dimensional data by mixtures of factor analyzers’, *Computational Statistics & Data Analysis*, 41(3–4), pp. 379–388.

Meldorf, M., Treufeldt, Ü. and Kilter, J. (2007) ‘Temperature dependency of electrical network load’, *Oil Shale*, 24(2), pp. 237–247.

Mets, K., Depuydt, F. and Develder, C. (2016) ‘Two-stage load pattern clustering using fast wavelet transformation’, *IEEE Transactions on Smart Grid*, 7(5), pp. 2250–2259.

Muscas, C., Pau, M., Pegoraro, P.A., *et al.* (2015) ‘Multiarea distribution system state estimation’, *IEEE Transactions on Instrumentation and Measurement*, 64(5), pp. 1140–1148.

Muscas, C., Pegoraro, P.A., Sulis, S., *et al.* (2016) ‘Fast multi-area approach for distribution system state estimation’, *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Taipei, Taiwan, pp. 1–6.

Mutanen, A. (2008) ‘*Sähkönjakeluverkon tilaestimoinnin täydentäminen kaukoluettavilla mittauksilla (Supplementing distribution network state estimation with remote measurements)*’, M.Sc. thesis, Tampere University of Technology, Finland. (in Finnish)

Mutanen, A. (2010) '*Customer classification and load profiling based on AMR measurements*', Tampere University of Technology, Research report.

Mutanen, A., Olmeda, D., Repo, S., *et al.* (2015) *'Deliverable 5.1 - State estimation and forecasting algorithms on MV & LV networks*', Ideal Grid for All (IDE4L), Project report.

Naka, S., Genji, T., Yura, T. and Fukuyama, Y. (2003) 'A hybrid particle swarm optimization for distribution state estimation', *IEEE Transactions on Power Systems*, 18(1), pp. 60–68.

Nanchian, S., Majumdar, A. and Pal, B.C. (2017) 'Three-phase state estimation using hybrid particle swarm optimization', *IEEE Transactions on Smart Grid*, 8(3), pp. 1035–1045.

Neimane, V. (2001) '*On development planning of electricity distribution networks*', Ph.D. thesis, KTH Royal Institute of Technology, Sweden.

Niknam, T. and Firouzi, B.B. (2009) 'A practical algorithm for distribution state estimation including renewable energy sources', *Renewable Energy*, 34(11), pp. 2309–2316.

Niska, H. (2013) 'Extracting controllable heating loads from aggregated smart meter data using clustering and predictive modelling', *8th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, Melbourne, Australia, pp. 368–373.

Norén, C. (1997) '*Typical load shapes for six categories of Swedish commercial buildings*', M.Sc. thesis, Lund Institute of Technology, Sweden.

Norén, C. and Pyrko, J. (1998a) 'Using multiple regression analysis to develop electricity consumption indicators for public schools', *ACEEE Summer Study on Energy Efficiency in Buildings*, Pacific Grove, CA, U.S., pp. 3.255–3.266.

Norén, C. and Pyrko, J. (1998b) 'Typical load shapes for Swedish schools and hotels', *Energy and Buildings*, 28(1), pp. 145–157.

Norén, C. and Pyrko, J. (1999) 'An analysis of electricity consumption and load demand in Swedish grocery stores', *ECEEE Summer Study on Energy Efficiency and CO2 Reduction: The Dimensions of the Social Challenge*, Mandelieu-la Napoule, France, pp. 1–13.

Nurmiranta, M. (2017) '*Data driven load modelling and customer behavior change detection*', M.Sc. thesis, Tampere University of Technology, Finland.

Nusrat, N., Irving, M. and Taylor, G. (2012) 'Novel meter placement algorithm for enhanced accuracy of distribution system state estimation', *IEEE Power and Energy Society General Meeting*, San Diego, CA, U.S., pp. 1–8.

Nusrat, N., Lopatka, P., Irving, M., *et al.* (2015) 'An overlapping zone-based state estimation method for distribution systems', *IEEE Transactions on Smart Grid*, 6(4), pp. 2126–2133.

Pau, M., Pegoraro, P.A. and Sulis, S. (2013) 'Efficient branch-current-based distribution system state estimation including synchronized measurements', *IEEE Transactions on Instrumentation and Measurement*, 62(9), pp. 2419–2429.

Pau, M., Ponci, F., Monti, A., *et al.* (2017) 'An efficient and accurate solution for distribution system state estimation with multiarea architecture', *IEEE Transactions on Instrumentation and Measurement*, 66(5), pp. 910–919.

PE EPS (2012) '*Functional requirements and technical specification of AMI/MDM system*', [Online], PE Electric Power Industry of Serbia (PE EPS), Available at: www.eps.rs/Eng/Documents/Functional requirements AMI MDM system version 3.0.pdf, (Accessed 1.12.2016).

Pereira, J., Saraiva, J.T. and Miranda, V. (2004) 'An integrated load allocation/state estimation approach for distribution networks', *8th International Conference on Probabilistic Methods Applied to Power Systems*, Ames, IA, U.S., pp. 180–185.

Pertl, M., Heussen, K., Gehrke, O. and Rezkalla, M. (2016) 'Voltage estimation in active distribution grids using neural networks', *IEEE Power and Energy Society General Meeting (PESGM)*, Boston, MA, U.S., pp. 1–5.

Piao, M., Shon, H.S., Lee, J.Y. and Ryu K.H. (2014) 'Subspace projection method based clustering analysis in load profiling', IEEE Transactions on Power Systems, 29(6), pp. 2628–2635.

Pitt, B.D. and Kirschen, D.S. (1999) 'Application of data mining techniques to load profiling', *21st International Conference on Power Industry Computer Applications (PICA)*, Santa Clara, CA, U.S., pp. 131–136.

Puromäki, A. (1959a) 'Sähkölaitoksen kuormituskäyräanalyysi moniregressio-menetelmällä (Analysis of electric utility's load curve with multiple regression method)', *Voima ja Valo*, 1959(5–6), pp. 116–123. (in Finnish)

Puromäki, A. (1959b) 'An experiment to analyse load curves with the multiple-regression method', Presented as an appendix in general report prepared by the load profiling committee, *International Union of Producers and Distributors of Electrical Energy (UNIPEDE) World Conference*, Lausanne, Switzerland, pp. 1–4.

Prahastono, I., King, D. and Ozveren, C.S. (2007) 'A review of electricity load profile classification methods', *42nd International* Universities Power Engineering Conference (UPEC), Brighton, U.K., pp. 1187–1191.

Primadianto, A., Lin, W.T., and Lu, C.N. (2016) 'Performance comparison of distribution system state estimation methods', *6th IEEE Innovative Smart Grid Technologies Conference Asia (ISGT-Asia)*, Melbourne, Australia, pp. 1121–1126.

Provoost, F. (2011) 'The use of smart meters to improve customer load models', *21st International Conference on Electricity Distribution (CIRED)*, Frankfurt, Germany, pp. 1–4.

REG 1.3.2009/66. Valtioneuvoston asetus sähköntoimitusten selvityksestä ja mittauksesta (The Finnish Council of State statue on settlement and measurement of electricity transactions). (in Finnish)

Rencher A.C. (2002) '*Methods of multivariate analysis'*, 2nd Edition, New York, NY, U.S., John Wiley & Sons Inc.

Repo, S., Della Giustina, D., Ravera, G., *et al*. (2011) 'Use case analysis of real time low voltage network management', *2nd IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Manchester, U.K., pp. 1–8.

Riihimäki, H. and Koponen, P. (2012) *'Prediction of energy consumption from outdoor temperature for houses electrically heated via heat storage*', VTT Technical Research Centre of Finland, Report number: VTT-R-02882-12.

Rousseeuw, P.J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.

Räsänen, T. and Kolehmainen, M. (2009) 'Feature-based clustering for electricity use time series data', *9th International Conference on Adaptive and Natural Computing Algorithms (ICANNGA)*', Kuopio, Finland, pp. 401–412.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K. and Kolehmainen, M. (2010) 'Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data', *Applied Energy*, 87, pp. 3538–3545.

Salvador, S. and Chan, P. (2004) 'Determining the number of clusters/segments in hierarchical cluster/segmentation algorithms', *16th IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, Florida, USA, pp. 576–584.

Sarić, A. and Ćirić, R. (2003) 'Integrated fuzzy state estimation and load flow analysis in distribution networks', *IEEE Transactions on Power Delivery*, 18(2), pp. 571–578.

Satopää, V., Albrecht, J., Irvin, D. and Raghavan, B. (2006) 'Finding a "kneedle" in a haystack: detecting knee points in system behaviour', *International Conference on Distributed Computing Systems Workshops (ICDCSW)*, Minneapolis, MN, USA, pp. 166–171.

Schiavo, L.L., Larzeni, S., Vailati, R., *et al.* (2015) 'Cost/benefit assessment for large-scale smart grid projects: the case of project of common interest for smart grid "GREEN-ME"', *23rd International Conference on Electricity Distribution (CIRED)*, Lyon, France, pp. 1–6.

Schwarz, G. (1978) 'Estimating the dimension of a model', *The Annals of Statistics*, 6(2), pp. 461–464.

Sener (1992) *'Sähkön käytön kuormitustutkimus 1992 (Research on electricity use 1992)*', Sener, Report number: 7103. (in Finnish)

Seppälä, A. (1996) '*Load research and load estimation in electricity distribution*'. Ph.D. thesis, Helsinki University of Technology, Finland.

Seppälä, A. (2007) '*Tyyppikäyrämenettelyn laskentaohje (Calculation instructions for customer class load profiling*)', [Online], Enease Oy, Available at: http:// energia.fi/files/553/Tyyppikayramenettelyn_laskentaohje.pdf, (Accessed 6.2.2017). (in Finnish)

Shafiu, A., Jenkins, N. and Strbac, G. (2005) 'Measurement location for state estimation of distribution networks with generation', *IEE Proceedings - Generation, Transmission and Distribution*, 152(2), pp. 240–246.

Siirto, O. (1989) '*Sähkön tarpeen lämpötilariippuvuuden tutkiminen (Research on electricity use temperature dependency)*', M.Sc. thesis, Helsinki University of Technology, Finland. (in Finnish)

Singh, R., Pal, B.C. and Jabr, R.A. (2010) 'Statistical representation of distribution system loads using Gaussian mixture model', *IEEE Transactions on Power Systems*, 25(1), pp. 29–37.

SLY (1986) '*Sähkön käytön kuormitusmittaukset (Measurements on electricity use)*', Suomen Sähkölaitosyhdistys r.y. (SLY), Report number: 1/1986. (in Finnish)

SLY (1992) '*Verkoston mitoitusenergiat (Network dimensioning energies)*', Suomen Sähkölaitosyhdistys r.y. (SLY), Report number: SA 10:92. (in Finnish)

Střelec, M, Janeček, P., Georgiev, D., Zápotocká, A. and Janeček, E. (2015) 'Backward/forward probabilistic network state estimation tool and its real world validation', *56th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON)*, Riga, Lithuania, pp. 1–6.

SULPU (2015) '*Lämpöpumpputilasto 2015 (Heat pump statistics 2015)*', [Online], Finnish Heat Pump Association (SULPU), Available at: www.sulpu.fi/documents/ 184029/208772/Lämpöpumpputilasto 2015, kuvaajat (1).pdf, (Accessed 28.12.2016). (in Finnish)

Tabachnick, B. and Fidell, L. (2006) *'Using Multivariate Statistics'*, 5th Edition, Needham Heights, MA, USA, Allyn & Bacon Inc.

Teng, J.-H. (2002) 'Using voltage measurements to improve the results of branch-current-based estimators for distribution systems', *IEE Proceedings - Generation, Transmission & Distribution*, 149(6), pp. 667–672.

Tibshirani, R., Walter, G. and Hastie, T. (2001) 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B*, 63(2), pp. 411–423.

Tsekouras, G.J., Hatziargyriou, N.D. and Dialynas, E.N. (2007) 'Two-stage pattern recognition of load curves for classification of electricity customers', *IEEE Transactions on Power Systems*, 22(3), pp. 1120–1128.

UNFCCC (2016) '*The Paris agreement*', [Online], United Nations Framework Convention on Climate Change (UNFCCC), Available at: http://unfccc.int/paris_agreement/items/9485.php, (Accessed 1.11.2016).

Vasudevan, K., Atla, C. S. R. and Balaraman, K. (2015) 'Improved state estimation by optimal placement of measurement devices in distribution system with DERs', *International Conference on Power and Advanced Control Engineering (ICPACE)*, Bangalore, India, pp. 253–257.

Vayá M.G., Koller M., Wyss V., *et al*. (2016) 'Demand response based on smart metering infrastructure to facilitate PV interaction in low voltage grids', *CIRED Workshop*, Helsinki, Finland, pp. 1–4.

Vercamer, D., Steurtewagen, B., Van del Poel, D. and Vermeulen, F. (2016) 'Predicting consumer load profiles using commercial and open data', *IEEE Transactions on Power Systems*, 31(5), pp. 3693–3701.

Verdú, S.V., Garcia, M.O., Franco, F.J.G., *et al*., (2004) 'Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters', *IEEE PES Power Systems Conference and* Exposition, New York, NY, U.S., pp. 899–906.

Wan, J. and Miu, K. (2003) 'Weighted least squares methods for load estimation in distribution networks', *IEEE Transactions on Power Systems*, 18(4), pp. 1338–1345.

Wang, H. and Schulz, N.N. (2004) ‘A revised branch current-based distribution system state estimation algorithm and meter placement impact’, *IEEE Transactions on Power Systems*, 19(1), pp. 207–213.

Warne, R. (2014) ‘A primer on multivariate analysis of variance (MANOVA) for behavioral scientists’, *Practical Assessment, Research & Evaluation*, 19(17), pp. 1–10.

Winkler, R., Klawonn, F. and Kruse, R. (2012) ‘Problems of fuzzy c-means clustering and similar algorithms with high dimensional data sets’, In: Gaul W., Geyer-Schulz, A., Schmidt-Thieme, L., Kunze, J. (eds) *Challenges at the Interface of Data Analysis, Computer Science, and Optimization - Studies in Classification, Data Analysis, and Knowledge Organization*, Berlin, Germany, Springer, pp. 79–87.

Wu, J., He, Y. and Jenkins, N. (2013) ‘A robust state estimator for medium voltage distribution networks’, *IEEE Transactions on Power Systems*, 28(2), pp. 1008–1016.

Wu, K-L. and Yang, M-S. (2005) ‘A cluster validity index for fuzzy clustering’, *Pattern Recognition Letters*, 26, pp. 1275–1291.

Xie, X.L. and Beni, G. (1991) ‘A validity measure for fuzzy clustering’, *IEEE Transactions on Pattern Analysis and Machine Learning*, 13(8), pp. 841–847.

Xygkis, T.C., Korres, G.N. and Manousakis, N.M. (2016) ‘Fisher information based meter placement in distribution grids via the D-optimal experimental design’, *IEEE Transactions on Smart Grid*, PP(99), pp.1–1 (Accepted for publication).

Zhang, T., Ramakrishnan, R. and Livny, M. (1996) ‘BIRCH: an efficient data clustering method for very large databases’, *ACM SIGMOD International Conference on Management of Data*, Montreal, Canada, pp. 103–114.

Zhao, Q. and Fränti, P. (2009) ‘Sum-of-squares based cluster validity index and significance analysis’. *International Conference on Adaptive and Natural Computing Algorithms (ICANNGA)*, Kuopio, Finland, pp. 313–322.

Publication 1

A. Mutanen, S. Repo, and P. Järventausta, "AMR in distribution network state estimation," presented at the 8th Nordic Electricity Distribution and Asset Management Conference (NORDAC), Bergen, Norway, Sept. 8–9, 2008.

AMR IN DISTRIBUTION NETWORK STATE ESTIMATION

Antti Mutanen[*)], *Sami Repo, Pertti Järventausta*
Tampere University of Technology, P.O. Box 692, FI-33101 Tampere, Finland
[*)] *e-mail antti.mutanen@tut.fi*

ABSTRACT

In the past few years, many distribution utilities have shown increasing interest towards distribution automation with the hope that automation will ultimately lead to a more efficient and economic operation of distribution networks. An important part of distribution automation is the real-time monitoring and control of distribution networks. Distribution automation functions such as network loading and voltage control, reactive power regulation, control of distributed generation and demand side management require accurate real-time estimates of network voltages and line flows. In this paper, the use of automatic meter reading (AMR) to improve the accuracy of distribution network state estimation is proposed. The efficient use of AMR measurements, especially the voltage measurements, is problematic in present distribution network state estimation systems. To fully utilize AMR measurements a new branch-current-based state estimation algorithm is introduced. Finally the benefits of using AMR measurements in state estimation are verified with MATLAB simulations.

1. INTRODUCTION

The goal of distribution state estimation (DSE) is to obtain the best possible estimate of the state of the network by processing the available information. Nowadays DSE relies mainly on the substation measurements, network data and load curves. Substation measurements include real time measurements from busbar voltages and feeder currents or powers. With these measurements it is possible to adjust the feeder loads accurately, but the load distribution inside the feeders remains uncertain. This uncertainty is mainly caused by inaccuracies in the load curves. The statistical mean values given in the load curves can differ from the true consumption. Since the load estimates are inaccurate also the line current and voltage level estimates inside the feeders are inaccurate.

Requirements for DSE accuracy have grown tighter. Customers have started to demand higher quality of supply and distribution utilities are adopting active network management methods in order to minimize the network investment costs. Correct voltage level is an integral part of electricity quality. The quality of supply is inadequate if the supply voltage is not within a specific range around the nominal voltage. The standard EN 50160 defines acceptable limits to the supply voltage variations. Under normal conditions during each period of one week 95 % of the 10 minute mean root-mean-square values of the supply voltage must be within ±10 % of the nominal voltage [1]. However, this is only the minimum requirement. In practice distribution companies often have more strict targets for the voltage quality.

To effectively control the distribution network voltage level network operators or automatic voltage control applications need accurate real-time estimates of network voltages. Also other active distribution network management functions such as control of distributed generation, reactive power regulation, feeder reconfiguration and restoration, and demand side management require accurate real-time estimates of network voltages and line flows. [2] Active network management is often used to increase the utilization degree of the existing

networks. Increasing the utilization degree helps to postpone network reinforcements but it also reduces the margins for acceptable network operation states. Accurate state estimation helps to monitor that the network operating state stays within these margins.

The simplest way to enhance DSE accuracy is to increase the number of real-time measurements. Power, current and voltage measurements along medium voltage (MV) feeders are an effective way to increase state estimation accuracy. Unfortunately, it is too expensive to add MV measurements to distribution networks solely for state estimation purposes. A perfect state estimate could be achieved by measuring all the loads, but practical constraints make it difficult. The investment costs often prevent distribution utilities from installing meters on every secondary substation and the data transmission issues prevent the real-time reading of all AMR meters. New methods are needed to minimize the amount and cost of measurements needed to improve the DSE accuracy. Another problem is that present DSE applications cannot use all available measurements. Especially the use of voltage measurements to improve the load estimation is impossible in the present applications. This paper solves these problems by using specially selected low voltage (LV) measurements and introducing a new branch-current-based state estimation algorithm.

The novelty of this paper is that AMR meters are used to enhance the DSE accuracy in a cost-effective way. AMR measurements are cheap to use since many distribution networks already have vast numbers of AMR meters. AMR meters have been installed primarily for the remote reading of electricity consumption, but their remote reading capabilities can also be used in state estimation. AMR meters are capable of measuring active power and voltage in real-time. Some meters can also measure reactive power and are capable of measuring power flows in both directions. This paper describes how AMR meters can be used to enhance DSE accuracy. First the new state estimation algorithm is presented and then the effect of AMR measurements on state estimation accuracy is demonstrated with MATLAB simulations.

2. STATE ESTIMATION

2.1 Basic DSE

DSE is a multi-stage process that combines information from many different sources and uses it to calculate the state of the network. Figure 1 presents the basic flow chart for DSE. State estimation starts with load estimation. Finnish DSE applications use load curves to estimate electricity consumption. The Finnish load research project has defined hourly load models for 46 different customer classes [3]. Each load curve gives the customer's average hourly loads and standard deviations for every hour of the year. The loads are given in active power but the load curves also include customer class specific power factors. Furthermore, the load curves define temperature correlation factors for each customer class so that the load estimates can be adjusted with outdoor temperature measurements.

At the second stage of DSE previously estimated loads and network information are used in load flow calculation. Present DSE programs use the backward/forward method to calculate the distribution network load flow. The load flow is calculated only for medium voltage network therefore the low voltage loads are summed to secondary substation connection points. The network topology and the line parameters are attained from the network information system. As a result of load flow calculation preliminary line flows and node voltages are acquired. This result can be referred as a first level state estimate. At this point the state estimate is still very inaccurate. The third stage employs distribution network measurements to improve the state estimation accuracy.

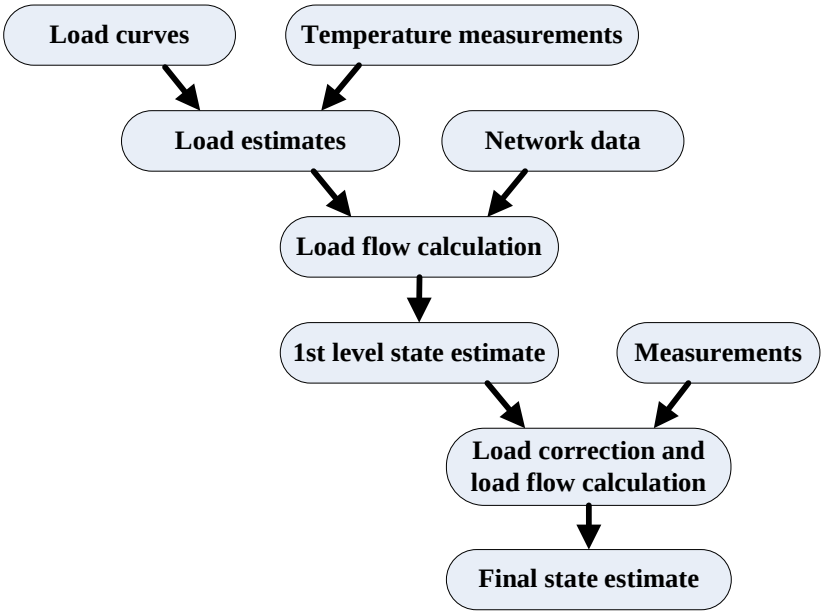


Figure 1. Basic flow chart for distribution state estimation.

In present DSE methods feeder line flow measurements are used to correct the load estimates so that the estimated line flows correspond to the measured line flows. The difference between estimated and measured feeder power flow is distributed to the load estimates in relation to their standard deviations. After the load estimates have been adjusted to fit the line flow measurements the load flow is recalculated and the final state estimate is acquired. This load correction method is simple and easy to implement. The downside is that only power measurements can be used to adjust load estimates. Current measurements need to be coupled with voltage measurements to produce power measurements. Independent voltage measurements can be used only to set initial voltages in forward sweep part of the load flow calculation. Use of the backward/forward method also limits the present DSE methods to the estimation of radial networks.

2.2 New DSE methods

Requirements for more accurate state estimation and future needs to estimate meshed distribution networks have led to the development of new DSE methods. The first new generation DSE methods were presented almost 15 years ago and since then many new DSE methods have been proposed. Most of these new methods are based on a weighted least squares (WLS) approach. WLS estimation has been used in transmission state estimation (TSE) since 1970s [4]. Applying TSE for distribution systems is a challenging task. The limited number of real-time measurements, high resistance to reactance ratios and current measurements cause problems for traditional TSE algorithms. Despite the difficulties, several studies have successfully applied the basics of TSE methods in DSE [5–7].

For distribution networks Baran and Kelley have proposed a branch-current-based DSE method [8]. It is also based on the WLS approach, but it uses branch currents as state variables where as the traditional TSE methods use node voltages as state variables. In later studies the branch-current-based method has been developed further. Its computation speed has been improved [9], possibility to use voltage measurements has been added [10] and state variables have been converted into a polar form [11].

When considering requirements for future DSE applications, the most important features are the accuracy and possibility to use all kind of real-time measurements. The effective use of real-time measurements requires that both current and voltage measurements

can be used independently in DSE. In this paper, a branch-current-based DSE algorithm is chosen to handle the AMR measurements. This choice helps to avoid the current measurement problems associated with TSE-based algorithms. The branch-current-based algorithms are also shown to be faster than the node-voltage-based algorithms [8;10].

3. PROPOSED STATE ESTIMATION ALGORITHM

3.1 Algorithm formulation

The state estimation algorithm used in this paper is based on the algorithm presented by Wang and Schulz. A detailed description of the algorithm can be found in the paper [11]. The algorithm uses the magnitudes and phase angles of the branch currents as state variables. The benefit of using current magnitudes as state variables is that current magnitude measurements correspond directly with state variables. The algorithm uses WLS estimation to determine the most likely state of the network. In WLS estimation the goal is to minimize the weighted sum of squared measurement residuals. Measurement residual is the difference between measured and estimated value and each residual is weighted with the variance of the corresponding measurement.

Some modifications were done to the original algorithm. The algorithm was altered to use equivalent single phase circuits and equality constraints were added to handle the zero-injection measurements. The use of equality constraints helped to avoid the ill-condition problems arising from the combination of high and low weights associated to zero-injection and pseudo load measurements. The equality constrained WLS problem can be solved by using the method of Lagrange multipliers [12]. In the method of Lagrange multipliers the constrained minimization problem is solved by minimizing the Lagrangian function

$$L(\boldsymbol{x},\boldsymbol{\lambda}) = \frac{1}{2}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})]^T\boldsymbol{R}^{-1}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})] + \boldsymbol{\lambda}^T\boldsymbol{c}(\boldsymbol{x}) \qquad (1)$$

where $\boldsymbol{x}$ is the state vector
$\boldsymbol{\lambda}$ is the Lagrange multiplier vector
$\boldsymbol{z}$ is the measurement vector
$\boldsymbol{h}(\boldsymbol{x})$ is the measurement function
$\boldsymbol{R}$ is the covariance matrix ($\boldsymbol{R} = \mathrm{diag}[\sigma_1^2 \quad \sigma_2^2 \quad \cdots \quad \sigma_N^2]$ where σ_i^2 is the variance of the measurement i)
$\boldsymbol{c}(\boldsymbol{x})$ is the zero-injection measurement function.

The minimization problem can be solved by differentiating $L(\boldsymbol{x},\boldsymbol{\lambda})$ partially with respect to $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ and setting the differentials to zero. This yields the following equations:

$$\frac{\partial L(\boldsymbol{x},\boldsymbol{\lambda})}{\partial \boldsymbol{x}} = -\boldsymbol{H}^T\boldsymbol{R}^{-1}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})] + \boldsymbol{C}(\boldsymbol{x})\boldsymbol{\lambda} = 0 \qquad (2)$$

$$\frac{\partial L(\boldsymbol{x},\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{c}(\boldsymbol{x}) = 0 \qquad (3)$$

where $\boldsymbol{H} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}$ and $\boldsymbol{C} = \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{x}}$ are the Jacobian matrices.

Equations (2) and (3) form a system of equations which can be solved iteratively by the Newton–Raphson method. At each iteration, the incremental change to the state vector ($\Delta \boldsymbol{x}$) is calculated with equation

$$\begin{bmatrix} \Delta \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H} & \boldsymbol{C}(\boldsymbol{x})^T \\ \boldsymbol{C}(\boldsymbol{x}) & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{H}^T \boldsymbol{R}^{-1} [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x})] \\ -\boldsymbol{c}(\boldsymbol{x}) \end{bmatrix}. \quad (4)$$

3.1 Algorithm steps

The proposed algorithm is composed of 7 basic steps seen in Figure 2. The algorithm calculates state estimates for one feeder at a time. The first two steps retrieve the network data and calculate the network load flow using the load estimates. The purpose of the load flow calculation is to obtain initial values for the state variables and node voltages. Good initial values improve the convergence characteristics of the Newton–Raphson algorithm. Substation voltage measurement is used in the load flow calculation to fix the voltage at the beginning of the feeder. Other real-time measurements are arranged into a measurement vector, which contains all measurements values. Variances that describe the measurements accuracies are gathered into the covariance matrix.

The iterative part of the algorithm starts with the calculation of the measurement function, zero-injection measurement function and corresponding Jacobian matrices. Equation (4) is used to calculate $\Delta \boldsymbol{x}$, which is then added to the state vector. Once the state vector has been updated the network voltages are recalculated using the forward sweep method. If the largest element of $\Delta \boldsymbol{x}$ is smaller than the convergence tolerance ε, then the state estimate is ready. Otherwise, another iteration cycle is performed.

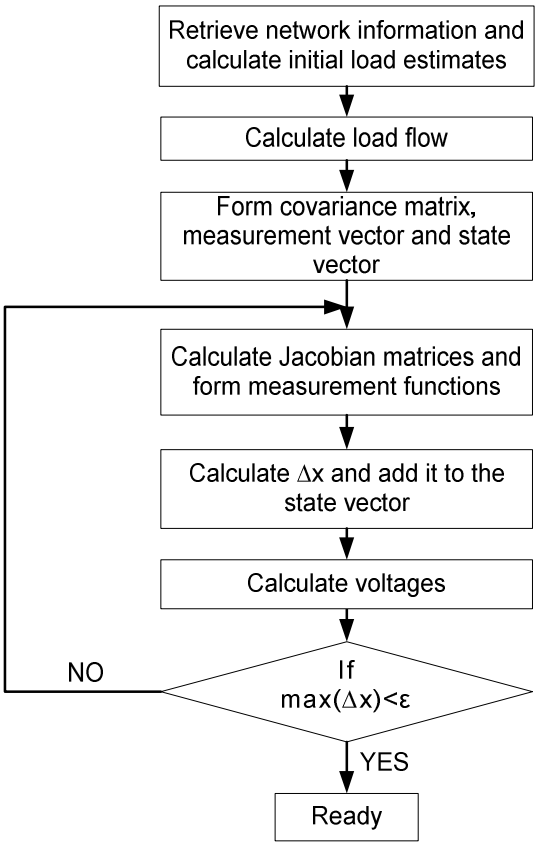


Figure 2. Basic steps for the proposed DSE algorithm.

4. SIMULATIONS

4.1. Test feeder

The above-presented algorithm was written into a MATLAB program, and its performance and the effect of using AMR measurements were tested with MATLAB simulations. IEEE 37-bus radial test feeder [13] was used in the case studies. The following modifications were made to the test feeder:

1) The voltage regulator was omitted.
2) The no-load transformer XFM-1 and the no-load node 775 were deleted.
3) All the loads were changed into constant PQ loads.
4) All the unsymmetrical loads were changed into symmetric three-phase loads and the feeder was modeled with an equivalent single phase circuit.
5) The nodes were renumbered for clarity.

The one-line diagram of the modified test feeder is shown in Figure 3.

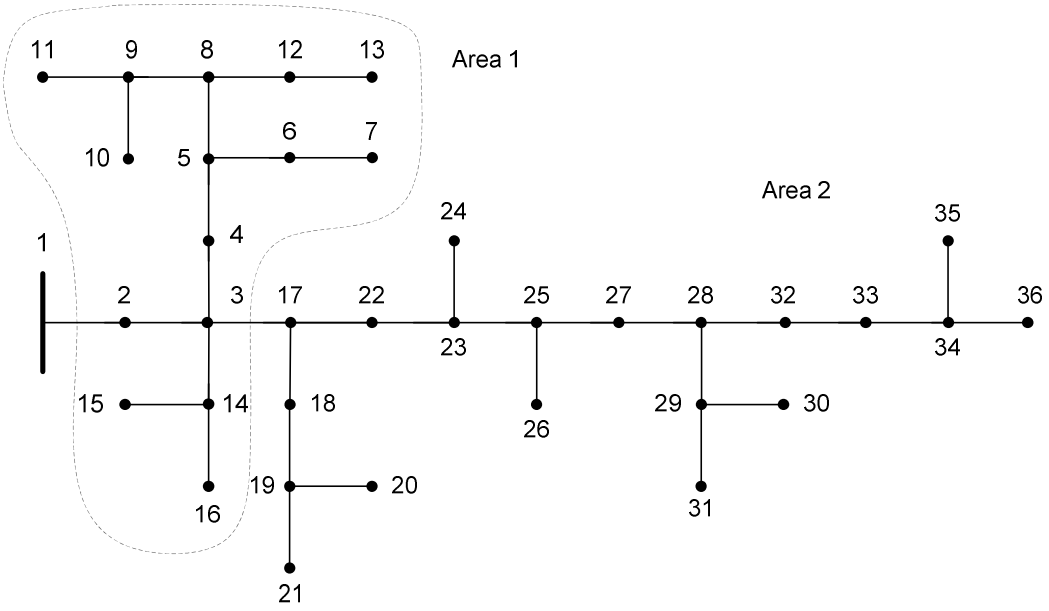


Figure 3. One-line diagram of the modified test feeder.

The test feeder was assumed to have a basic set of measurements: active and reactive power measurements at the beginning of the feeder, a voltage measurement at the node number 1 and pseudo measurements at all the load nodes. The power measurement accuracies were ±1 % and the voltage measurement was assumed to be ideal. The test feeder was divided into two areas where the pseudo MV load measurements had different accuracies. The pseudo measurements had a relative standard deviation (RSD) of 50 % in the area 1 and 20 % in the area 2. The above measurement configuration was referred as base case.

The simulations were performed first by varying the loads normally according to the pseudo measurement standard deviations. Then the true state of the feeder was calculated using the load flow program of Power System Toolbox [14]. The real-time measurements were created from the true states by varying them normally according to the measurement accuracy. Accuracy for the additional power measurements was ±1 % and ±0.2 % for the voltage measurements (95 % confidence level). The DSE was calculated with different measurement configurations and estimation errors were calculated for node voltages and branch currents by comparing the estimates with the true values. This procedure was repeated 1000 times for every measurement configuration and average errors were calculated.

4.2. Case 1: secondary substation measurements

In this case, secondary substation measurements were used to improve DSE accuracy. The measurements were installed on the low voltage side of the distribution transformers. Placing the measurements on the LV side of the distribution transformers is an economical solution because the LV measurements are much cheaper than the MV measurements. Using low voltage measurements requires modeling of the distribution transformers. In the test simulations 4.8/0.208 kV distribution transformers were added to the feeder model when necessary. The rating of the transformers was 200 kVA, except the one connected to the node 2, which was 1000 kVA. The transformers were presumed to have a relative short-circuit resistance of 1.15 % and reactance of 3.8 %.

The simulations showed that secondary substation power measurements are an effective way to enhance the DSE accuracy. For example, measuring the active and reactive powers from the nodes 2, 11 and 32 halved the errors associated to the voltage estimates and decreased the errors in the line current estimates by 34 %. The measurement location had a significant effect on the DSE accuracy. The best locations were found on the load nodes that had large standard deviations i.e. large loads that had inaccurate load estimates. It was also beneficial for the loads to be located far from the substation. Figure 4 shows how the voltage estimation errors decreased when the number of power measurements was increased. The benefits of additional measurements diminished after the best measurement locations had been occupied.

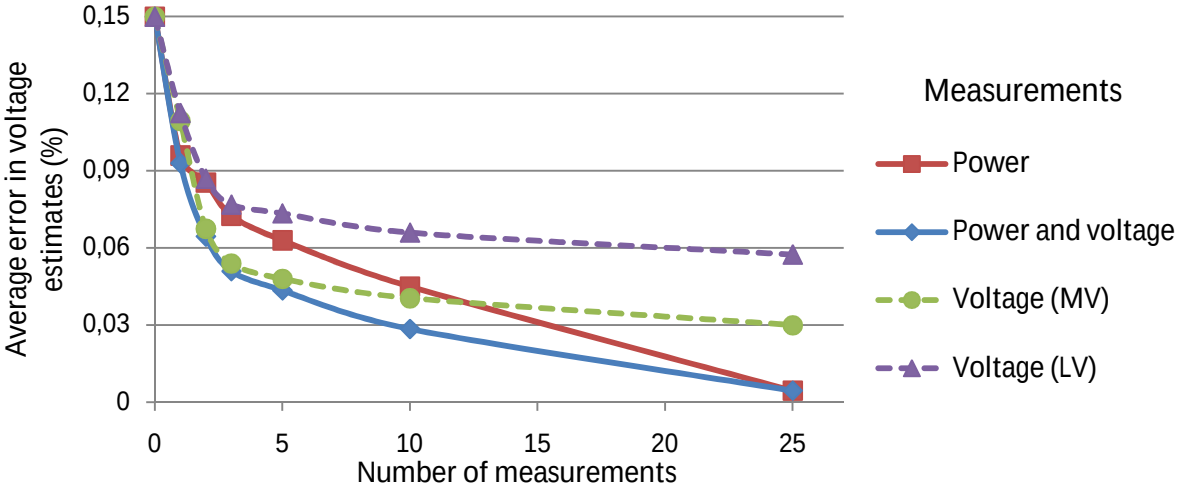


Figure 4. The effect of secondary substation measurements on the voltage estimation accuracy.

Power measurements were effective and they can be used in present DSE systems. Even better results were achieved by combining the power measurements with voltage measurements and using the proposed DSE algorithm. The combined power and voltage measurements allowed us to double the voltage estimation accuracy with only two measurement points. The solid lines in Figure 4 show that the combined power and voltage measurements were more accurate than the power measurements alone.

The voltage measurements can also be used separately from the power measurements. With a low number of measurements, the voltage measurements were almost as accurate as the combined power and voltage measurements. The dashed lines in Figure 4 show that the best results were attained when the voltages were measured from the medium voltage side of the distribution transformers. When the voltages were measured from the low voltage side, the unknown voltage drop in the transformer reduced the state estimation accuracy.

Case 2: AMR measurements in low voltage network

To fully utilize AMR measurements from low voltage networks, the LV networks need to be modeled to the DSE system. The LV networks shown in Figures 5 a) and b) were added to the test feeder model. In area 1 the LV loads consisted of four identical commercial loads and in area 2 there were 25 identical residential loads. The load sizes were set so that the sum of the LV loads matches with the corresponding MV loads. All the pseudo LV load measurements had a RSD of 100 %. The commercial loads were connected to the distribution transformer with an own LV feeder while in residential networks there were several loads connected along the feeders. Each LV line section between the network nodes had a resistance of 8.2 mΩ and reactance of 4.9 mΩ.

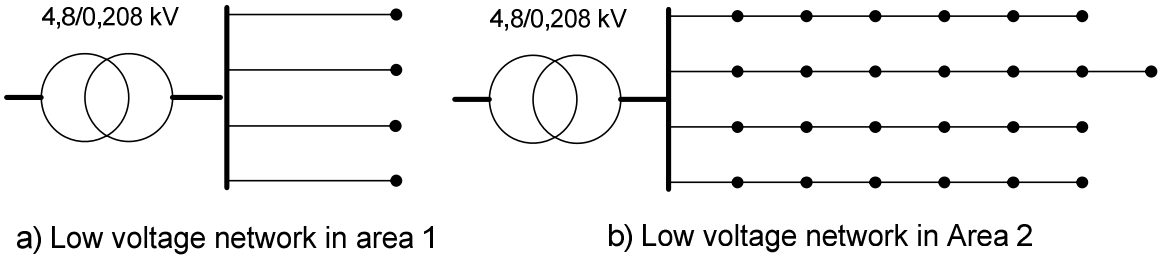


Figure 5. The low voltage networks used in the simulations.

Small amount of power measurements from the LV loads can enhance the DSE accuracy significantly only if the LV networks contain large loads. Figure 6 a) shows that measuring the 10 largest LV loads from the area 1 increased the voltage estimation accuracy by 50 %. The same measurements also improved the line current estimates by 34 %. If the LV network contains only small loads, it is difficult to enhance the DSE accuracy with a reasonable amount of power measurements. Measuring the 10 biggest residential loads from the area 2 improved the voltage estimation accuracy only by 4 %.

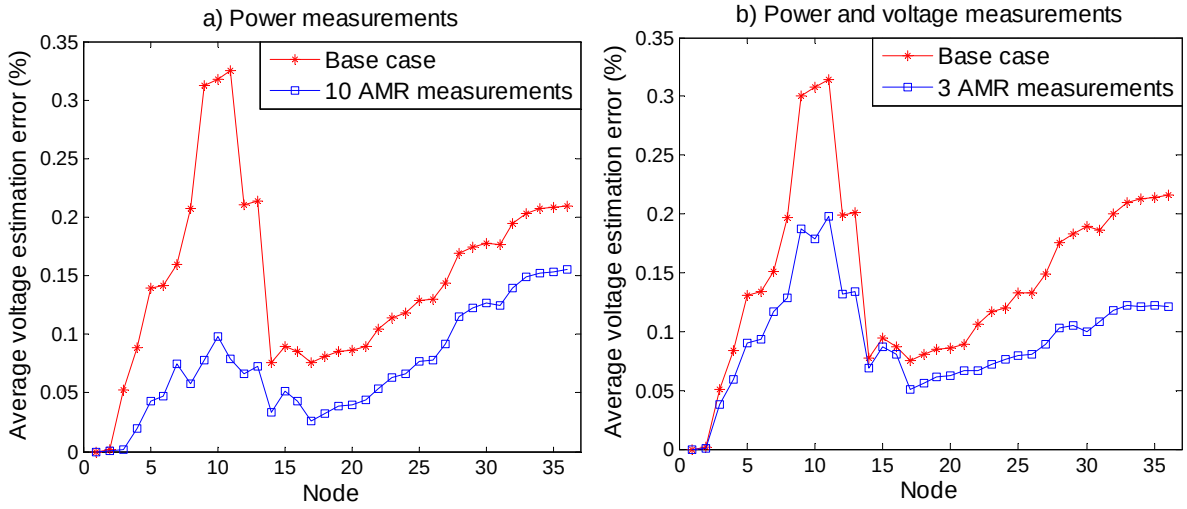


Figure 6. The effect of AMR measurements on the voltage estimation accuracy.

Measuring power from a single residential customer had virtually no effect on the DSE accuracy. As the secondary substation measurements also AMR meters can be utilized more effectively if power measurements are coupled with voltage measurements. At best, measuring power and voltage from a single residential customer in area 2 improved the MV network voltage estimates 18 % and current estimates 10 %. When the amount of AMR

measurements was increased to three, the corresponding improvements were 38 % and 19 %. Figure 6 b) shows the improvements in voltage estimates when one AMR measurement was located in area 1 and two in area 2. Even the accuracy of the voltage measurements was reduced by the unknown voltage drop in the distribution transformer and in LV line, the combined power and voltage measurements provided superior results compared to the power measurements alone.

The best locations for combined power and voltage measurements were at the end of the MV feeder branches and behind lightly loaded distribution transformers. Inside the LV network the ideal location was on a large customer connected to the distribution transformer with an own LV feeder. Good results were achieved also by measuring the first customer on a multi-customer feeder. Measuring the last customer provides little improvement to the MV state estimates, but increases the estimation accuracy on the LV feeder.

5. DISCUSSION

Good measurement accuracy is important when voltage measurements are used to improve DSE accuracy. Figure 7 a) shows how the estimation errors increased when the voltage measurement accuracy was reduced. The figure is based on three voltage measurements from MV network nodes 11, 30 and 36. Voltage sensors used in MV measurements usually achieve a measurement accuracy of ±0.5–1.0 %. Better accuracy requires more expensive voltage transformers. Residential AMR meters also have a voltage measurement accuracy of ±0.5–1.0 %. Only some industrial AMR meters reach ±0.2 % measurement accuracy. The voltage measurement accuracy must be improved to fully exploit the potential of the proposed DSE method.

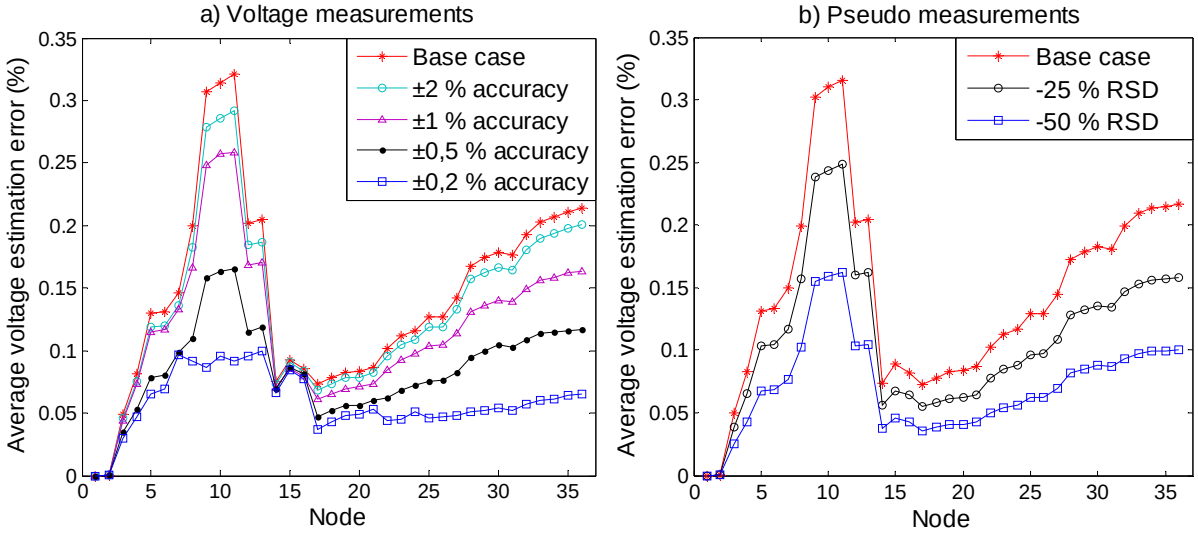


Figure 7. The DSE accuracy a) with different voltage measurement accuracies and b) with improved pseudo measurement accuracies.

The use of voltage measurements clearly reduced the number of measurements needed to achieve substantial improvements. The test feeder contained some very large commercial loads and therefore even a low number of power measurements provided good results. In a purely residential network, the benefits of using combined power and voltage measurements would have been even clearer.

Since it is not economically viable to read every single AMR meter in real-time, it is important to achieve the desired level of estimation accuracy with a reasonable amount of

measurements. The real-time reading of all AMR meters is expensive and requires a considerable amount of data transmission capacity. Near real-time reading, for example once every hour, requires less data transmission capacity, but the measurement delays decrease the accuracy of the DSE.

Other possibilities to use AMR meters should also be studied. AMR meters are capable of recording load profiles in short time intervals. These load profiles can be stored to the meter and read during normal meter reading without increasing the meter reading frequency. With the help of the load profile data, the classification of customers can be reviewed and more specific customer classes or even individual load curves can be formed. Since the load curves are used as pseudo measurements, improving the load curve accuracy affects the estimation accuracy. Figure 7 b) shows how the voltage estimation accuracy would improve if the pseudo measurement accuracy was increased 25 % or 50 %.

6. CONCLUSIONS

Voltage magnitude measurements contain a lot of information about the network states, but to extract this information a new type of state estimation methods are needed. This paper presented a branch-current-based DSE method that can use all available measurement types, including voltage measurements, to enhance the accuracy of node voltage and line flow estimates. The proposed method is based on a WLS approach and uses branch currents as state variables.

Simulations with the proposed DSE method showed that the use of voltage measurements can reduce the number of metering points needed to achieve accurate state estimates. Especially the benefit of low voltage AMR measurements was increased considerably. With the help of the new DSE method and AMR measurements, the future DSE accuracy requirements can be satisfied cost-effectively.

REFERENCES

1. EN 50160:1999 E. Voltage characteristics of electricity supplied by public distribution systems. Brussels 1999, CENELEC. 14 p.
2. Strbac, G., Jenkins, N., Hird, M., Djapic, P. & Nicholson, G. Integration of operation of embedded generation and distribution networks. Manchester 2002, Manchester Centre for Electrical Energy, Final Report K/EL/00262/REP URP 02/1145, 94 p.
3. Seppälä, A. Load research and load estimation in electricity distribution. Dissertation. Espoo 1996, VTT Publications 298. 118 p. +app. 19 p.
4. Schweppe, F.C. & Handschin E.J. Static State Estimation in Electric Power Systems. Proceedings of the IEEE, Vol. 62, No. 7, July 1974. pp. 972–982.
5. Baran, M. & Kelley, A. State Estimation for Real-Time Monitoring of Distribution Systems. IEEE Transaction on Power Systems, Vol. 9, No. 3, August 1994. pp. 1601–1609.
6. Wan, J. & Miu, K. Weighted Least Squares Methods for Load Estimation in Distribution Networks. IEEE Transactions on Power Systems, Vol. 18, No. 4, November 2003. pp. 1338–1345.
7. Cobelo, I., Shafiu, A., Jenkins, N. & Strbac, G. State estimation of networks with distributed generation. European Transactions on Electrical Power, Vol. 17. No. 1, January/February 2007. pp 21-36.

8. Baran, M. & Kelley, A. A Branch-Current-Based State Estimation Method for Distribution Systems. IEEE Transactions on Power Systems, Vol. 10, No. 1, February 1995. pp. 483–491.
9. Lin, W.-M., Teng, J.-H. & Chen, S.-J. A Highly efficient Algorithm in Treating Current Measurements for Branch-Current-Based Distribution State estimation. IEEE Transactions on Power Delivery, Vol. 16, No. 3, July 2001. pp. 433–439.
10. Teng, J.-H. Using voltage measurements to improve the results of branch-current-based estimators for distribution systems. IEE Proceedings - Generation, Transmission & Distribution, Vol. 149, Issue 6, November 2002. pp. 667–672.
11. Wang, H. & Schulz, N.N. A Revised Branch Current-Based Distribution System State Estimation Algorithm and Meter Placement Impact, IEEE Transactions on Power Systems, Vol. 19, No. 1, February 2004. pp. 207-213.
12. Wu, F.F., Liu, W.E. & Lun, S.M. Observability Analysis and Bad Data Processing for State Estimation with Equality Constraints. IEEE Transactions on Power Systems, Vol. 3, No. 2, May 1988. pp. 541-548.
13. Radial Test Feeders - IEEE Distribution System Analysis Subcommittee. [Online] Available: http://www.ewh.ieee.org/soc/pes/dsacom/testfeeders.html.
14. Power System Toolbox - Cherry Tree Scientific Software. [Online] Available: http://www.eagle.ca/~cherry/pst.htm.

Publication 2

A. Mutanen, A. Koto, A. Kulmala, and P. Järventausta, "Development and testing of a branch current based distribution system state estimator," presented at the 46th International Universities' Power Engineering Conference (UPEC). Soest, Germany, Sep. 5–8, 2011.

Available at: http://ieeexplore.ieee.org/document/6125578

Development and Testing of a Branch Current Based Distribution System State Estimator

Antti Mutanen
Tampere University of Technology, Finland
antti.mutanen@tut.fi

Antti Koto
Tampere University of Technology, Finland
antti.koto@tut.fi

Anna Kulmala
Tampere University of Technology, Finland
anna.kulmala@tut.fi

Pertti Järventausta
Tampere University of Technology, Finland
pertti.jarventausta@tut.fi

***Abstract*- The recent increase of distributed generation has forced many distribution network operators to develop distribution automation and active network management. Many active distribution network management functions need accurate real-time estimates of the network state. In this paper, a distribution network state estimation algorithm is developed and used in conjunction with coordinated voltage control. The state estimator utilizes equality constrained weighted least squares optimization and includes bad data detection. The state estimator is tested with MATLAB simulations, real-time digital simulator and in a real distribution network.**

***Index Terms* -- Bad data detection, distribution system state estimation, equality constraints, testing, weighted least squares.**

I. Introduction

The purpose of distribution system state estimation (DSSE) is to obtain the best possible estimate of the network state by processing available information. Nowadays DSSE relies mainly on substation measurements, network data and load profiles. The substation measurements include real-time measurements of busbar voltages and feeder current or power flows. With these measurements it is possible to adjust the feeder loads accurately, but the load distribution inside the feeders remains uncertain.

There is a need for more accurate DSSE because the amount of distribution automation and active control is constantly increasing. Active distribution network management functions such as voltage level management, control of distributed generation, reactive power regulation, feeder reconfiguration and restoration, and demand side management require accurate real-time estimates of network voltages and line flows. Especially the increase of distributed generation is an important driver for state estimation development [1].

DSSE can be made more accurate by adding measurements to the distribution network and using advanced state estimation methods. In the last 15 years, several new DSSE methods have been proposed in the literature [1]–[4]. Many of them are based on the weighted least squares method, but the selection of state variables varies. Some are using node voltages [1], [2] as state variables whereas others have chosen to use branch currents [3], [4].

In order to make DSSE more accurate, we developed a branch current based distribution system state estimator exploiting equality constrained weighted least squares optimization [5]. The state estimator was formulated to utilize all real-time current, power and voltage measurements available in a distribution network. The developed state estimator was written into a MATLAB program, and its performance and the effect of the additional current, power and voltage measurements were tested with MATLAB simulations.

In this paper, the state estimator is further developed by adding bad data detection using state estimation residuals. Furthermore, the state estimator is coupled with a coordinated voltage control algorithm [6] and tested in a Real-Time Digital Simulator (RTDS) and in a real distribution network.

This paper will first revise the formulation of the developed DSSE method and introduce the used bad data detection method. Thereafter, test results from MATLAB and RTDS simulation and real-life demonstration are presented. The test results are discussed and conclusions are given at the end.

II. Formulation

A. Main algorithm

The state estimation algorithm in this paper is based on the method presented by Wang and Schulz [4]. The algorithm uses the magnitudes and phase angles of branch currents as state variables. The benefit of using current magnitudes as state variables is that current magnitude measurements, which are the dominating measurement types in distribution systems, correspond directly with state variables. The algorithm uses weighted least squares (WLS) estimation to determine the most likely state of the network. In WLS estimation the goal is to minimize the weighted sum of squared measurement residuals. Measurement residual is the difference between measured and estimated value and each residual is weighted with the variance (accuracy) of the corresponding measurement.

Some modifications were done to the original algorithm. The algorithm was altered to use equivalent single phase circuits and equality constraints were added to handle the zero-injection measurements. The use of equality constraints helped to avoid the ill-condition problems arising from the combination of high and low weights associated to zero-injection and pseudo load measurements. The equality constrained WLS problem can be solved by using the method of Lagrange multipliers [7]. In the method of Lagrange multipliers the constrained minimization problem is solved by minimizing the Lagrangian function

$$L(\boldsymbol{x},\boldsymbol{\lambda}) = \frac{1}{2}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})]^T\boldsymbol{R}^{-1}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})] + \boldsymbol{\lambda}^T\boldsymbol{c}(\boldsymbol{x}), \quad (1)$$

where $\boldsymbol{x}$ is the state vector
$\boldsymbol{\lambda}$ is the Lagrange multiplier vector
$\boldsymbol{z}$ is the measurement vector
$\boldsymbol{h}(\boldsymbol{x})$ is the measurement function
$\boldsymbol{R}$ is the measurement covariance matrix ($\boldsymbol{R} = \mathrm{diag}[\sigma_1^2 \quad \sigma_2^2 \quad \cdots \quad \sigma_N^2]$ where σ_i^2 is the variance of the measurement i)
$\boldsymbol{c}(\boldsymbol{x})$ is the zero-injection measurement function.

The minimization problem can be solved by differentiating $L(\boldsymbol{x},\boldsymbol{\lambda})$ partially with respect to $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ and setting the differentials to zero. This yields the following equations:

$$\frac{\partial L(\boldsymbol{x},\boldsymbol{\lambda})}{\partial \boldsymbol{x}} = -\boldsymbol{H}^T\boldsymbol{R}^{-1}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})] + \boldsymbol{C}(\boldsymbol{x})\boldsymbol{\lambda} = 0 \quad (2)$$

$$\frac{\partial L(\boldsymbol{x},\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \boldsymbol{c}(\boldsymbol{x}) = 0 \quad (3)$$

where $\boldsymbol{H} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}$ and $\boldsymbol{C} = \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{x}}$ are the Jacobian matrices.

Equations (2) and (3) form a system of equations which can be solved iteratively by the Newton–Raphson method. At each iteration, the incremental change to the state vector ($\Delta\boldsymbol{x}$) is calculated with equation

$$\begin{bmatrix}\Delta\boldsymbol{x}\\ \boldsymbol{\lambda}\end{bmatrix} = \begin{bmatrix}\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H} & \boldsymbol{C}(\boldsymbol{x})^T\\ \boldsymbol{C}(\boldsymbol{x}) & 0\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{H}^T\boldsymbol{R}^{-1}[\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})]\\ -\boldsymbol{c}(\boldsymbol{x})\end{bmatrix}. \quad (4)$$

Once the calculation has converged, the node voltages can be determined with a forward sweep calculation.

B. Bad data detection

Bad data detection is an essential part of any state estimator. State estimators must be able to detect, identify and remove bad data from the measurement set. Measurements may contain errors due to various reasons. Meters can have biases, drifts or wrong connections. Telecommunication system failures can also lead to large deviations in recorded measurements.

Some measurement errors are easy to detect with simple logical rules. For example, negative voltage and current magnitudes and measurements, which are several orders of magnitude larger or smaller than expected, are easily recognized as bad data. Unfortunately, not all types of bad data are detected that easily. However, in more indistinct cases, other detection methods can be utilized.

In WLS state estimation, the bad data detection can be made by examining the measurement residuals. This has to be done after the estimation process. The bad data detection is essentially based on the statistical properties of the residuals. One of the most used bad data detection methods is the *Largest Normalized Residual* r_{max}^N -test. This test is composed of the following steps [8]:

1. Solve the WLS estimation and obtain the elements of the measurement residual vector ($\boldsymbol{r}$):

$$\boldsymbol{r} = \boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x}) \quad (5)$$

2. Compute the normalized residuals ($\boldsymbol{r}^N$):

$$\boldsymbol{r}^N = \frac{|\boldsymbol{r}|}{\sqrt{\boldsymbol{\Omega}_{\mathbf{ii}}}}, \quad (6)$$

where $\boldsymbol{\Omega}_{\mathbf{ii}}$ is $diag(\boldsymbol{\Omega})$
$\boldsymbol{\Omega}$ is $Cov(\boldsymbol{r})$.

3. Find the largest normalized residual ($\boldsymbol{r}_{max}^{\boldsymbol{N}}$).
4. If $\boldsymbol{r}_{max}^N > c$, then the corresponding measurement is erroneous. Here, c is a chosen detection threshold, for instance 3.0.
5. If bad data is detected, eliminate the faulty measurement from the measurement set and go back to step 1.

The faulty measurements are eliminated one by one. After each elimination, WLS state estimation procedure is repeated.

The largest normalized residual test can detect bad data if the removal of the corresponding measurement does not render the system unobservable. It is possible to identify all cases of single bad data where the faulty measurements are not critical or belong to a critical pair or critical k-tuple. Critical measurements are those measurements whose removal would cause the system to become unobservable. A critical pair and k-tuple contain two or more measurements, respectively, whose simultaneous removal would make the system unobservable.

In the case of multiple bad data, only part of the measurements errors can be identified. Faulty measurements with weakly correlated measurement residuals can be identified. If the measurement residuals are strongly correlated, the bad data can be identified only in the case of non-conforming bad data. If the identification of faulty measurement fails, the largest normalized residual test can incorrectly remove a faultless measurement.

Because our state estimator is based on equality constrained WLS estimation, the measurement residual covariance matrix can not be solved as usual. Solution for this problem can be found from [7]. In equality constrained state estimation the measurement residual covariance matrix $\boldsymbol{\Omega}$ is equal to

$$Cov(\boldsymbol{r}) = \boldsymbol{R}^{-1} - \boldsymbol{H}\boldsymbol{E}_1\boldsymbol{H}^T, \quad (7)$$

where $\boldsymbol{E}_1$ is the upper left corner of the inverse of $\boldsymbol{F}$.

$$\boldsymbol{F}^{-1} = \begin{bmatrix}\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H} & \boldsymbol{C}(\boldsymbol{x})^T\\ \boldsymbol{C}(\boldsymbol{x}) & 0\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{E}_1 & \boldsymbol{E}_2{}^T\\ \boldsymbol{E}_2 & \boldsymbol{E}_3\end{bmatrix}, \quad (8)$$

where $\boldsymbol{C}(\boldsymbol{x})$ is the Jacobian matrix of the equality constraint function.

The problem with measurement residual based bad data detection is that it requires a certain amount of redundancy from the measurement configuration. In distribution networks, the number of measurements and thus also the redundancy level is very limited. In this paper, load models are used as load pseudo-measurements. With these artificial measurements it is possible to detect and identify rough errors in real measurements.

III. Testing

A. MATLAB simulations

The above-presented algorithm was written into a MATLAB program, and its performance was tested with MATLAB simulations. IEEE 37-bus radial test feeder [9] was used in the simulations. The following modifications were made to the test feeder:

1) The voltage regulator was omitted.
2) All the loads were changed into constant PQ loads.
3) All the unsymmetrical loads were changed into symmetric three-phase loads and the feeder was modelled with an equivalent single phase circuit. This is a common simplification in Finnish distribution network calculation.
4) The nodes were renumbered for clarity.

The one-line diagram of the modified test feeder is shown in Fig 1.

The test feeder was assumed to have a basic set of measurements: active and reactive power flow measurements at the beginning of the feeder, a voltage measurement at the node 1 and pseudo-measurements at the load nodes. The measurement accuracies were set to ±1 % for the power flow measurements and ±0.2 % for the voltage measurement (with a 95 % confidence level). The pseudo-measurements were given a relative standard deviation of 50 % in the area 1 and 20 % in the area 2.

Simulations comparing the proposed and existing Finnish DSSE methods were conducted. In the existing DSSE methods [10] only feeder line flow measurements are used to correct the load estimates and the difference between the estimated and the measured feeder power flow is distributed to the load estimates in relation to their standard deviations. The existing DSSE method was also modelled into the MATLAB.

The simulations were performed by first varying the loads normally according to the pseudo-measurement standard deviations. Then the true state of the feeder was calculated using a load flow program. The power flow and voltage measurements were created from the true states by varying them normally according to the corresponding measurement accuracy. Finally, the state estimates were computed and the estimation errors were calculated for node voltages by comparing the estimates with the true values. This procedure was repeated 10000 times and average errors were calculated.

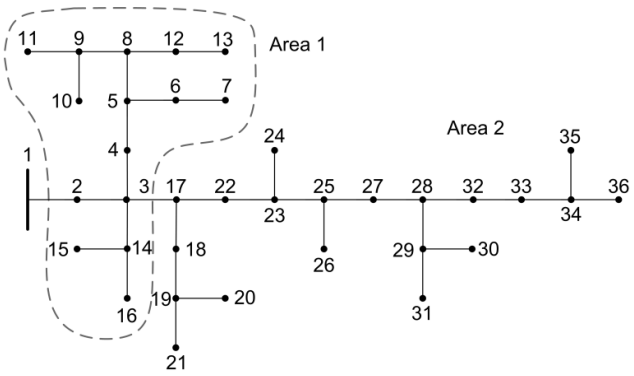


Fig. 1. One-line diagram of the modified test feeder.

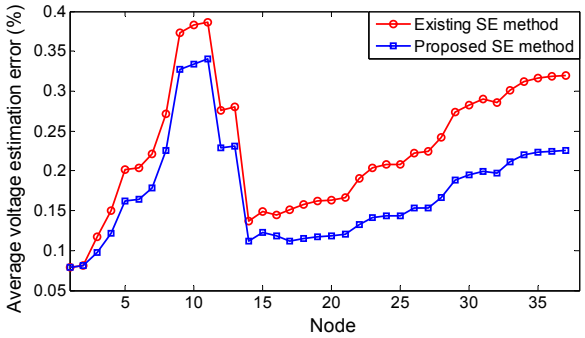


Fig. 2. Estimation accuracy comparison.

The proposed state estimation method provided 24 % smaller average voltage estimation error. The difference is shown in Fig. 2. Simulations were also done with additional power, current and voltage measurements to study their effects on the state estimation accuracy. Some of these results are published in [5].

B. RTDS simulations

In the next testing phase, the state estimation algorithm was coupled with a coordinated voltage control algorithm [6] and the resulting MATLAB prototype software was tested in RTDS environment. The purpose of RTDS simulations was to verify the correct operation of the prototype software before it is demonstrated in a real distribution network.

The coordinated voltage control algorithm aims to keep the network voltages between acceptable limits by controlling available active resources. In these simulations, it controls substation voltage and DG reactive power by changing the set points of substation automatic voltage control relay and DG automatic voltage regulator. The coordinated voltage control algorithm uses the results of the state estimation as inputs. In this paper, we concentrate on the state estimation part of the RTDS simulations. Simulation results from the coordinated voltage control point of view can be found in [11].

The simulation arrangement

In these simulations, RTDS is used to emulate a real distribution network. The simulation arrangement is depicted in Fig. 3. RTDS consists of hardware and software. The

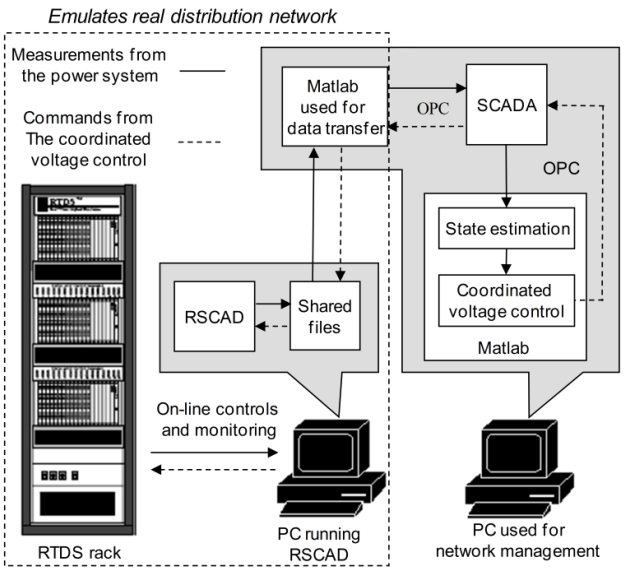


Fig. 3. RTDS simulation arrangement.

hardware is used to solve power system equations in real-time and is installed in a rack. The RSCAD software is run on an external computer and is used to construct the power system models and to control the simulations. The simulated network is controlled using commercial SCADA software (ABB MicroSCADA Pro SYS 600) and the prototype software containing both state estimation and coordinated voltage control algorithms. Measurement signals from the simulated network are transferred to SCADA. Data transfer between RSCAD and SCADA is realized using shared files.

The simulation network

The simulation network is constructed to correspond to the network in the forthcoming real-life demonstration. The network consists of two medium voltage feeders and contains one relatively large hydro power plant. The RTDS simulations are done with a three-phase network model. A reduced version of the real network model is used because of RTDS limitations. A single-phase representation of the simulation network is shown in Fig. 4.

The simulation network includes active and reactive power flow measurements at the beginning of each feeder and at the hydro power plant. Voltages are measured from the substation and from the power plant. The power plant breaker status is also monitored. Loads are modelled as symmetrical static constant power loads. In state estimation, the load pseudo-measurements are given a 10 % relative standard deviation. The distribution lines are modelled in both RSCAD and in the state estimator using a nominal π-model.

Simulation results

First, the state estimation results where compared to the monitored values in RSCAD to verify the accuracy of load flow calculation embedded in to the state estimator. When given ideal error-free measurements as inputs, the differences

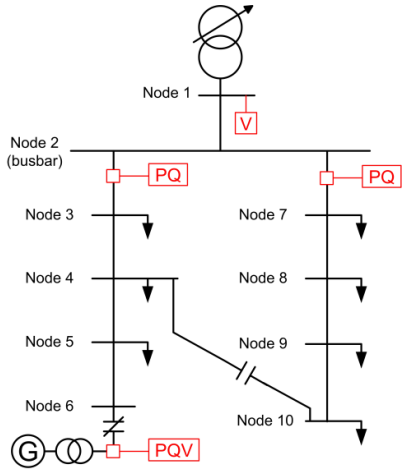


Fig. 4. RTDS simulation network.

in estimated and monitored node voltages were smaller than 0.01 %.

During the first set of RTDS simulation we noticed that the state estimation did not always converge when given highly conflicting inputs. Conflicting inputs can be caused, for example, by input synchronization errors. Synchronization errors were detected also in the RTDS simulations. Sometimes, when the power plant was disconnected from the network by opening the power plant breaker, the changed breaker status information reached the state estimator before the feeder power flow measurements had changed to correspond to the new topology. Bad data detection was added to the state estimator to tackle this problem.

The simulation network has a very low measurement redundancy, therefore detecting bad data is difficult. Only errors in power plant voltage measurement can be detected and identified directly from the measurement residuals. The feeder power flow measurements form critical pairs with the feeder load pseudo-measurements and with power plant power flow measurements. These groups of critical pairs are denoted here as *critical groups*. Removal of any measurement in a critical group would make the rest of the measurements critical. All the measurements in a critical group have equal normalized residuals, hence the erroneous measurement can not be identified. In order to identify faulty feeder power flow measurements we have to assume that the measurement errors can not be located in the load pseudo-measurements or power plant power flow measurements.

For example, if the load feeder reactive power flow measurement (Q_{27}) is erroneous, then that measurement and all reactive load pseudo-measurement on the same feeder (Q_7–Q_{10}) have identical normalized residuals. This is shown in Table 1. In order to identify the bad data, we have to assume that only the highlighted measurements in Table I can contain errors.

In RTDS simulations, the bad data detection threshold was set to 3.0, which is a typical bad data detection threshold in transmission system state estimation. With this threshold

value, the measurement Q_{27} can vary between 2.32 and 3.39 p.u. without being suspected as bad data. In Table I, the measurement Q_{27} is outside this range, its normalized residual is larger than 3.0 and it is identified as bad data.

In the case of the previously mentioned input synchronization error, the bad data detection fails because the state estimation does not converge. This problem was solved by first detecting the existence of bad data from the non-convergence and then running the state estimation again without the feeder power flow measurements. After the new pseudo-measurement based state estimate was calculated, the normalized residuals were calculated for the feeder power flow measurements. Measurement with the largest normalized residual was identified as bad data and removed from the measurement set. Then the state estimation was run again. This procedure was repeated until the state estimator converged and all erroneous measurements were removed. The power plant power flow and voltage measurements were removed from the measurements set based on the power plant breaker status. Table I shows the normalized measurement residuals in the case of the input synchronization error.

After adding the bad data detection, no further problems were encountered in RTDS simulations. The state estimator worked as planned supplying correct state estimates to the coordinated voltage control algorithm.

TABLE I
EXAMPLES OF NORMALIZED MEASUREMENT RESIDUALS DURING BAD DATA DETECTION

Mea-sure-ment	Erroneous Q_{27}			Input synchronization error		
	$\boldsymbol{z_{real}}$	$\boldsymbol{z}$	$\boldsymbol{r^N}$	$\boldsymbol{z_{real}}$	$\boldsymbol{z}$	$\boldsymbol{r^N}$
P_{23}	–1.35	–1.35	0.00	**8.56**	**–1.35**	**14.29**
P_{27}	11.26	11.26	0.17	11.26	11.26	0
Q_{23}	4.17	4.17	0.00	**2.09**	**4.17**	**10.89**
Q_{27}	**2.85**	**3.85**	**5.54**	2.85	2.85	0
V_6	1.05	1.05	0.00	-	-	-
P_2	0	0	0	0	0	0.00
P_3	6.80	6.80	0.00	6.80	6.80	0.00
P_4	1.15	1.15	0.00	1.15	1.15	0.00
P_5	0.60	0.60	0.00	0.60	0.60	0.00
P_6	–10.00	–10.00	0.00	0	0	0.00
P_7	3.06	3.06	0.07	3.06	3.06	0.00
P_8	4.93	4.93	0.00	4.93	4.93	0.00
P_9	1.94	1.94	0.00	1.94	1.94	0.00
P_{10}	1.12	1.12	0.01	1.12	1.12	0.00
Q_2	0	0	0	0	0	0.00
Q_3	1.89	1.89	0.00	1.89	1.89	0.00
Q_4	0.33	0.33	0.00	0.33	0.33	0.00
Q_5	0.17	0.17	0.00	0.17	0.17	0.00
Q_6	2.00	2.00	0.00	0	0	0.00
Q_7	0.88	0.88	**5.54**	0.88	0.88	0.00
Q_8	1.41	1.41	**5.54**	1.41	1.41	0.00
Q_9	0.55	0.55	**5.54**	0.55	0.55	0.01
Q_{10}	0.32	0.32	**5.54**	0.32	0.32	0.04

C. *Real-life demonstrations*

The operation of the previously presented prototype software was demonstrated in a real Finnish distribution network in May 2010. The demonstration arrangement was somewhat similar to the one shown in Fig. 3. The parts inside the dashed line were replaced by the real distribution network and the prototype software was run on a PC separate to the network management PC running SCADA and DMS. As a safety feature, the operator executed the control commands from the coordinated voltage algorithm manually. As in the Fig. 4, the demonstration network consisted of two medium voltage feeders and one power plant. Instead of the power flow measurements, only current flow measurements were available from the beginning of the feeders. The network and loads were modelled into the MATLAB with the same detail as in the network information system. The load models included hourly load estimates and their standard deviations for each distribution transformer.

During the demonstration of the coordinated voltage control algorithm, some problems were detected in the state estimation. The bad data detection identified the feeder current flow measurements incorrectly as bad data. This was caused by the exceptionally warm weather during the demonstration. The average daily temperature was over 10 °C higher than normally in May. The probability of such weather occurring in May is less than 3 %. High temperature caused a radical drop in heating loads and the bad data detection interpreted low feeder current flows as faulty measurements. This problem could have been avoided if the load temperature dependencies had been taken into account. The state estimator included a load temperature correction feature, but no temperature dependencies were available for the used load models. The bad data detection had to be turned off. Further problems were experienced because of an inaccurate substation voltage measurement. The used voltage measurement had a measurement resolution of 1 % (0.2 kV). This reduced the voltage estimation accuracy significantly. Despite these problems, the coordinated voltage control demonstration was completed successfully [12].

Next, we aimed to verify the results in Fig. 2 by comparing the developed state estimator and the state estimator in a real distribution management system (ABB MicroSCADA Pro DMS 600). Inputs and outputs from the DMS state estimator were saved for later off-line comparison. This required some special arrangements because the DMS 600 does not normally save the state estimates. The DMS 600 source code was edited to save the state estimation results into a database. The state estimation results were then read to the MATLAB through ODBC interface. Finally, the state estimation results and inputs; feeder current flows and substation voltage measurements, were written into a text file. The state estimation was run once an hour and the results were saved for a period of one week. Data was collected from one medium voltage feeder. To find out the true voltages, two voltage measurements were added to distribution

transformers at branch ends of the studied feeder. These measurements were done with AMR meters with power quality monitoring functions.

After one week, the data collection PC was retrieved from the distribution network operator's control room and the results were analyzed. We noticed that the DMS state estimator had not corrected the loads to match the feeder current flows. This was caused by human error; the current measurements were not connected to the network model in NIS. Secondly, we discovered that the DMS state estimator had used a different substation voltage measurement than assumed. Thereby, the results of the developed state estimator and DMS state estimator were not comparable.

Even without the above mentioned mistakes, the comparison of state estimators would have been difficult. The demonstration network was lightly loaded during the demonstration and the voltage drops were very small. The differences between estimated and measured node voltages would have been close to the voltage measurement accuracy. The state estimation accuracy could have been verified also by comparing the estimated and measured loads in the distribution transformers. Unfortunately, measurement of the transformer loads was not possible.

IV. Discussion and conclusions

The branch current based distribution system state estimator was further developed by adding bad data detection. Distribution networks have a very low measurement redundancy, thereby detecting bad data is difficult. In many cases, the only way to identify faulty measurements is to use pseudo-measurements in the bad data detection process. The real-life demonstration of the developed DSSE method proved that the commonly used 3σ bad data detection threshold is inadequate when using load profile data to detect errors in feeder line flow measurements. The bad data detection threshold should be raised and more accurate load models with temperature dependency correction should be used. Further research is needed to find out if these actions are enough to make the bad data detection work. Next, we are going to use AMR measurements to improve the accuracy of load models. After this we can retest the state estimation algorithm.

MATLAB simulations proved that the developed DSSE method is more accurate than the existing Finnish state estimation method. Demonstrating this improvement in a real distribution network is difficult. To see the differences in the voltage estimates, the demonstration network should have big stochastic loads, large voltage drops and very accurate voltage measurements.

Acknowledgements

The authors would like to thank ABB Distribution Automation unit for technical support and Koillis-Satakunnan Sähkö Oy for providing the real-life demonstration environment.

References

[1] I. Cobelo, A. Shafiu, N. Jenkins and G. Strbac, "State Estimation of Networks with Distributed Generation," European Transactions on Electrical Power, vol. 17, no. 1, pp 21–36, January/February 2007.

[2] C. Gouveia, H. Leite and I. Ferreira, "Voltage and Current Sensor for State estimation in Distribution Network with Generation," International Conference on Renewable Energies and Power Quality (ICREPO'10), Granada, Spain, 23–25 March, 2010.

[3] M. Baran and A. Kelley, "A Branch-Current-Based State Estimation Method for Distribution Systems," IEEE Transactions on Power Systems, vol. 10, no. 1, pp. 483–491, February 1995.

[4] H. Wang and N.N. Schulz, "A Revised Branch Current-Based Distribution System State Estimation Algorithm and Meter Placement Impact," IEEE Transactions on Power Systems, vol. 19, no. 1, pp. 207–213, February 2004.

[5] A. Mutanen, S. Repo, and P. Järventausta, "AMR in Distribution Network State Estimation," 8th Nordic Electricity Distribution and Asset Management Conference, Bergen, Norway, 8–9 September, 2008.

[6] A. Kulmala, S. Repo and P. Järventausta, "Increasing Penetration of Distributed Generation in Existing Distribution Networks Using Coordinated Voltage Control," International Journal of Distributed Energy Resources, vol. 5, no. 3, pp 227–255, 2009.

[7] F.F. Wu, W-H.E. Liu and S-H Lun, "Observability Analysis and Bad Data Processing for State Estimation with Equality Constraints," IEEE Transactions on Power Systems, vol. 3, no. 2, pp. 541–548, May 1988.

[8] A. Abur and A.G. Expósito, *Power System State Estimation: Theory and Implementation*. New York: Marcel Dekker, Inc. 2004.

[9] Radial Test Feeders - IEEE PES Distribution System Analysis Subcommittee. [Online]. Available: http://www.ewh.ieee.org/soc/pes/dsacom/testfeeders/index.html.

[10] M. Kärenlampi, P. Verho, P. Järventausta and J. Partanen, "Forecasting the Short-Term Loads of MV-Feeders and Distribution Substations - a DMS Function," Proceedings of 12th Power System Computation Conference, Dresden, Germany, 19-23 August, 1996, pp. 757–763.

[11] A. Kulmala, A. Mutanen, A. Koto, S. Repo and P. Järventausta, "RTDS Verification of a Coordinated Voltage Control Implementation for Distribution Networks with Distributed Generation," IEEE PES Innovative Smart Grid Technologies Conference Europe, Gothenburg, Sweden, 11–13 October, 2010.

[12] A. Kulmala, A. Mutanen, A. Koto, S. Repo and P. Järventausta, "Demonstrating Coordinated Voltage Control in a Real Distribution Network," unpublished.

Publication 3

A. Mutanen, S. Repo, P. Järventausta, A. Löf, and D.D. Giustina, “Testing low voltage network state estimation in RTDS environment,” presented at the 4th IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Copenhagen, Denmark, Oct. 6–9, 2013.

Available at: https://doi.org/10.1109/ISGTEurope.2013.6695482

Testing Low Voltage Network State Estimation in RTDS Environment

Antti Mutanen, Sami Repo and Pertti Järventausta
Department of Electrical Engineering
Tampere University of Technology
Tampere, Finland
antti.mutanen@tut.fi

Atte Löf
VTT Technical Research Centre of Finland
Espoo, Finland
Atte.Lof@vtt.fi

Davide Della Giustina
A2A Reti Elettriche Spa
Brescia, Italy
davide.dellagiustina@a2a.eu

***Abstract*— The low voltage network operating environment is going through changes. The simultaneous introduction of intermittent renewable energy production and customer requirements for increased power quality and supply reliability are forcing utilities to rethink the role of low voltage networks. With recent advances in smart grid technology, low voltage network automation is emerging as a viable option to traditional network investments. Congestion management and demand response, for example, can be used to keep the network currents and voltages within acceptable limits. In order to control the network, we must first have a comprehensive view on the state of the network. In this paper, the low voltage network monitoring concept proposed by the FP7 European project INTEGRIS is tested. Real-Time Digital Simulator (RTDS) is used to test how well the measurements from secondary substations and smart meters can be combined in a state estimator to get a real-time view of the network state.**

***Index Terms*—state estimation, low voltage, RTDS, smart grids**

I. INTRODUCTION

With the advent of smart grids, the ways of operating distribution networks are changing. The amount of distributed generation (DG) is increasing and in order to accommodate the intermittent DG with reasonable network investments, the automatic control of networks is increased. For example, demand response is introduced to keep the currents and voltages within acceptable limits. This is true for both medium voltage (MV) and low voltage (LV) networks. In LV networks, requirements for better power quality and distribution reliability and simultaneous increase in customer level DG are calling for novel automation solutions. Secondary substation automation, smart meters, demand response and home automation have been proposed as a solution. In INTEGRIS project, the above mentioned technologies are combined in order to fulfill the first mentioned requirements. The INTEGRIS project is part of the EU 7th Framework Program. The INTEGRIS project proposes a decentralized distribution network automation concept that can completely and efficiently fulfill the requirements of the smart grid networks of the future. The INTERIS concept includes both MV and LV levels. In this paper, only LV level is considered.

This research was funded by the EU 7th Framework Program project INTEGRIS.

Important part of the INTEGRIS project is the efficient utilization of measurement devices. Information from smart meters and secondary substations can be utilized in power quality management, fault management, monitoring and control of the network. The aim of this paper is to study how measurements from smart meters and secondary substations can be utilized in LV network monitoring. In this paper, these measurements are combined in a state estimator in order to get the best possible view of the LV network state. The proposed LV network monitoring concept is presented and tested in a Real-Time Digital Simulator (RTDS). The power quality aspect of INTEGRIS concept is studied in [1].

In order to control the distribution network, it is essential to know the state of the network i.e. node voltage and line current magnitudes. Voltage information is needed for example in network voltage/VAR control and current information in network congestion management. Accurate state estimation enables the automation of distribution network control. In this paper, the accuracy of LV network state estimation is evaluated with different measurement configurations, meter reading frequencies and measurement averaging times. Furthermore, the effect of load profiles on the state estimation accuracy is discussed and load profiling results from the INTEGRIS demonstration in Italy are shown.

II. LV NETWORK MONITORING CONCEPT AND RTDS LABORATORY SETUP

In INTEGRIS concept, distribution network automation is based on decentralized intelligence. Low voltage network monitoring and management is done on secondary substation (SS) level. Measurements from remote terminal units (RTUs), smart meters and home energy management systems (HEMSs) are collected, stored and analyzed in a single device called INTEGRIS Device (I-Dev) located at the SS. LV network state estimation is performed in I-Dev based on the available measurements and LV network model. If the state of the network is not acceptable, control commands are sent to smart meters and HEMSs. Customer level automation then decides how to control distributed energy recourses and controllable loads. Only processed information such as alarms, requests and aggregated data are sent to the higher level automation systems located in primary substations and control centre. [2]

Testing was made with RTDS equipment designed to simulate electrical power systems and test physical equipment in real-time. RTDS hardware is used in conjunction with RSCAD software that contains the network model. Single-phase presentation of the LV network modelled in RSCAD is shown in Fig 1. The test network is relatively small due to RTDS node limit. Still, the model is realistic as the test network is part of a real three-phase LV network in Finland. Although not shown in Fig. 1, the simulation model included also models for supplying network (voltage source and impedance) and secondary transformer. The test network contained seven three-phase residential customers, located at nodes 3, 5, 7, 9, 10, 12 and 13. The load on each customer node was based on real load measured from corresponding residential customer nodes. The load in node 13 was based on a measurement done with one second measurement interval while the other loads were based on ten minute measurement interval. These measurements were averaged or interpolated to corresponding ten second values and were updated into the RSCAD every tenth second. Shorter updating interval was not possible since changing all the load values in RSCAD took approximately seven seconds.

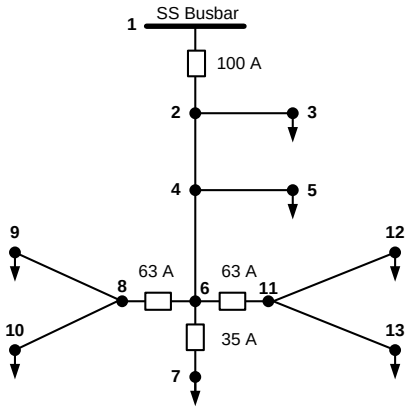


Figure 1. Single-phase presentation of the LV network used in the RTDS-simulations.

Schneider Electric's Easergy Flair 200C substation monitoring unit served as a RTU monitoring device and was set to measure current at the beginning of the LV feeder and voltage on the secondary substation LV busbar. The currents were measured using wireless sensors that used ZigBee communication protocol. Easergy Flair 200C measures voltage with ±1 % accuracy. Kamstrup 382L and Indra Emiel EBM-M65A smart meters were used to measure load on the customer nodes. These meters have class B and A (by EN 50470) energy measurement accuracies, respectively. The three-phase voltage and current values from RTDS were sent to RTU and smart meters via amplifiers that boosted the voltage and current signals to level which correspond the real values. With only two amplifiers available, each simulation had to be repeated several times in order to get measurement values from every customer.

Fig. 2 shows how the devices and components relevant to this state estimation study connect to the RTDS. In addition to these devices, Theregate home energy management device and MX Electrix power quality monitoring unit were also connected to the RTDS. According to the INTEGRIS communication concept [2], the measurements were relayed to SS I-Dev by using the IEC 61850 protocol. A protocol gateway was used to translate the RTU communication into IEC 61850 compliant format. The smart meters used DLMS (Device Language Message Specification) protocol to communicate with the meter data concentrator. State estimation results, measurements and alarms were saved on SS I-Dev database and from there the network state information was sent to iGrid's iControl SCADA.

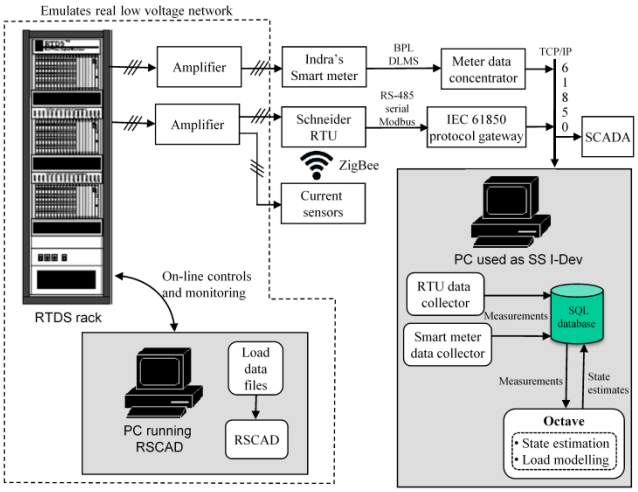


Figure 2. Connections in RTDS

III. State Estimation

The goal of distribution network state estimation is to obtain the best possible estimate of the state of the network by processing the available information. In this case, the available information is network topology, line parameters, RTU and smart meter measurements and load profiles which are used as pseudo measurements. The state of the network is described by the node voltages or line currents. Transmission network state estimation has been studied since the 1970's and is now considered a routine task [3]. Though the technology for transmission system state estimation is mature, there have been only a few papers in literature addressing the problem of LV network state estimation [4], [5]. However, there are countless papers on MV network state estimation [6]–[10]. The problems of LV and MV network state estimation are closely related. Both, when compared with traditional transmission network state estimation, exhibit similar characteristics:

- radial topology
- high R/X ratio
- asymmetric loads
- low measurement redundancy
- current measurements.

In LV networks, the high variability of the loads makes the state estimation task even more challenging. Since the number of measurements in distribution networks is low, it is important to place these few measurements correctly. The meter placement problem has been studied in many papers [5]–[7]. In this paper, we have also studied how the meter reading frequencies and measurement averaging times affect on the state estimation accuracy.

State estimation is commonly based on the weighted least squares (WLS) method [4]. In WLS estimation the goal is to minimize the weighted sum of squared measurement residuals. Measurement residual is the difference between measured and estimated value and each residual is weighted with the variance (accuracy) of the corresponding measurement. The state of the network can be defined either with node voltage magnitudes and their phase angles or with line current magnitudes and their phase angles. The traditional transmission system state estimation uses node voltages as state variables. The node voltage based state estimation has been successfully applied to LV networks [4], [5] but the branch current based state estimation has been developed specifically for distribution networks. In this paper, we have used a three-phase branch current based state estimator exploiting equality constrained WLS optimization. The use of equality constraints helps to avoid the ill-condition problems arising from the combination of high and low weights associated to zero-injection and pseudo load measurements. For mathematical details, the reader is referred to [3] and [8].

The distribution network is not observable unless we have measurements from every customer node. It is not always possible to measure every customer point (or communicate to every meter). Therefore, we have used customer class load profiles as pseudo measurements. A customer class load profile contains load expectation and standard deviation values for every hour of the year for certain type of a customer. In this study, the most descriptive load profile for each customer was selected from a set of 46 customer class load profiles.

IV. RTDS-SIMULATION RESULTS

A. Communication Delays

The RTDS environment was used to test the communication delays of the proposed INTEGRIS communication architecture. The average time delay to get measurements from RTU unit was 1.8 seconds. The delay was calculated from the time when the request to get the measurements from RTU was send to the time when the measurements were received to the I-Dev database. The average time delay to get measurements from Indra's smart meter depends on the number of the modules installed. In this case, with three modules, the average time delay was 2 minutes and 20 seconds.

B. Accuracy Metrics for Monitoring and State Estimation

RTDS-simulations were run for a testing period of one day and average root mean square errors (ARMSE) were calculated for the monitored and estimated quantities using (1).

$$ARMSE = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(q_{mon/est}(t) - q_{real}(t)\right)^2}, \quad (1)$$

where $q_{mon/est}$ is the monitored or estimated quantity
q_{real} is the real instantaneous value from RSCAD
T is the testing period
m is the number of points of interest.

Only customer connection point voltage and current magnitudes were used in (1) when studying voltage and current monitoring and state estimation accuracies.

C. LV Network monitoring

The effects of meter reading frequency and averaging time (measurement averaging window) on the LV network monitoring accuracy were tested by comparing the real network states with the measured values available in I-Dev database. Fig. 3 shows how these two factors affect the average RMS error of the monitored current values, when real and reactive powers from all customers are measured. As expected, the average RMS error is lower when the reading frequency is higher and when the averaging time is shorter. State estimation was not used at this point. Therefore, Fig. 3 shows how accurately the state of the network can be monitored in the INTEGRIS concept without state estimation.

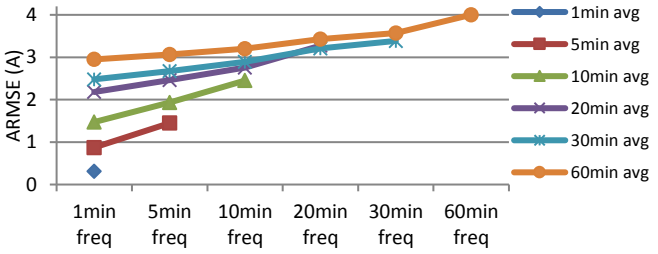


Figure 3. Average RMS error of monitored current values.

D. LV Network State Estimation

When state estimation was tested, the state estimator was given the following three-phase measurements as inputs:

- distribution substation LV busbar voltage
- LV feeder current
- real and reactive power measurements from all customer nodes.

The same meter reading frequencies and averaging times were used for both RTU and smart meter measurements. Fig. 4 and 5 show how the meter reading frequency and averaging times affect on the state estimation accuracy. 10 minute averaging time was used in Fig. 4 and one minute meter reading frequency was used in Fig. 5. The y-axis unit is either volt [V] or ampere [A] depending on the variable. Examining these figures, we can conclude that good state estimation accuracy is achieved only with high meter reading frequency and with short measurements averaging time. In practice, the optimization starts with the meter reading frequency, which is chosen as high as possible taking technical and economic constraints into account, then the measurement averaging time is set to the same value. If the measurement averaging time is set lower than the meter reading frequency, some of the measurement information will be lost. From Fig. 3–5 we can see that the estimated values are more accurate than the values monitored with smart meters only. By combining the low latency RTU measurements with the smart meter measurements, the state estimator is able to improve the customer node monitoring accuracy.

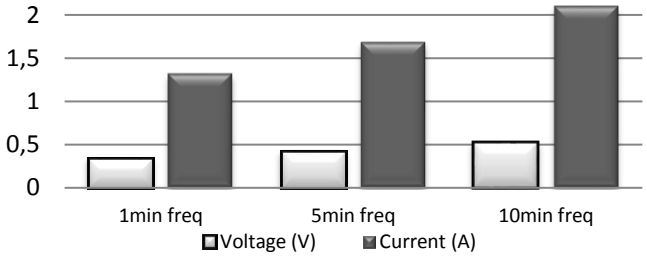


Figure 4. Average RMS error of estimated quantities when meter reading frequency is varied.

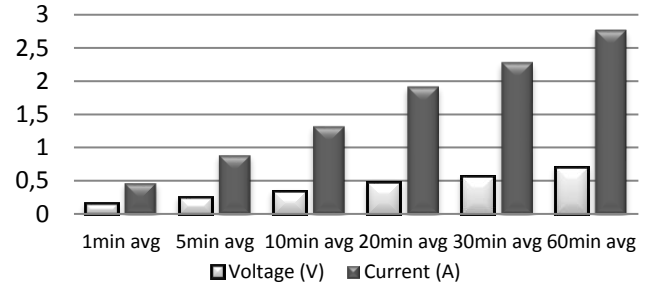


Figure 5. Average RMS error of estimated quantities when measurement averaging time is varied.

In smart grids, distribution network overloading situations can be avoided with automatic congestion management. When congestion management is applied, it is essential to estimate the network peak loads accurately. The peak load estimation accuracy depends largely on the measurement averaging time. Fig. 6 shows how accurately peak current on customer 10 is estimated when 10 and 30 minute averaging times are used. Clearly, the 30 minute averaging time is too long for accurately estimating the magnitude of current peaks. With 30 minute averaging time, the estimated peak current is over 20 % lower than the actual peak current. If the measurement averaging time is constant and the meter reading frequency is varied, the meter reading frequency will not have effect on the magnitude of estimated peak current, but it will affect how quickly the peak current is observed. The peak current study confirms that the meter reading interval and measurement averaging time should be selected equal and as short as possible. In a decentralized automation system, like the INTEGRIS concept, it is possible to set these parameters smaller than in traditional centralized automation. Within the LV network, the data transfer distances are short and the communication between smart meters and SS I-Dev can be implemented with broadband power line communication (BPL) without an external telecommunication company.

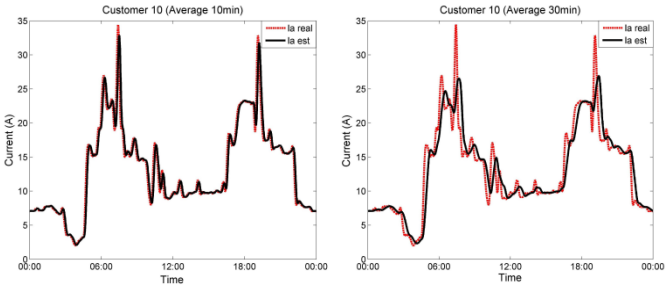


Figure 6. Estimated currents with 10 and 30 minute averaging times.

In previous simulations, it was assumed that all smart meters are available. However, sometimes there may be communication problems or meter malfunctions so that some or all smart meters are unavailable. Fig. 7 shows how the state estimation accuracy depends on the available smart meter configuration. Fig. 7 is based on simulations done with ideal smart meter data (real time measurements with zero delay) and is intended only for comparing different smart meter configurations. The best and the worst meter configurations are shown for cases; only one available smart meter, two available smart meters and one unavailable smart meter. The results show that the state estimation accuracy remains good even if one smart meter is offline but in the case of several unavailable smart meters the accuracy deteriorates significantly. When only a few smart meters were available, the best results were achieved when they were located on customers with high loads. In the case of few available smart meters, the state estimation accuracy could have been improved by utilizing customer node voltage measurements [5], [7].

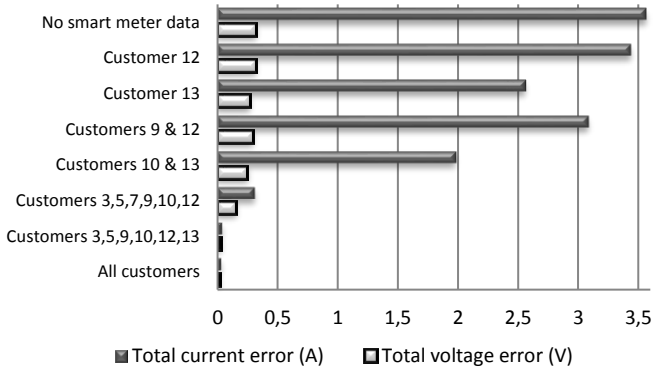


Figure 7. Average RMS error of estimated quantities when smart meter configuration is varied.

With a small number of available smart meters, the state estimation accuracy depends largely on the load pseudo measurement (load profile) accuracy. In this study, the load profiles assumed the loads to be symmetric even though the actual three-phase loads can be asymmetric. Fig. 8 shows an example of load asymmetry. Phase specific load profiles could be one way to improve the load profiling accuracy.

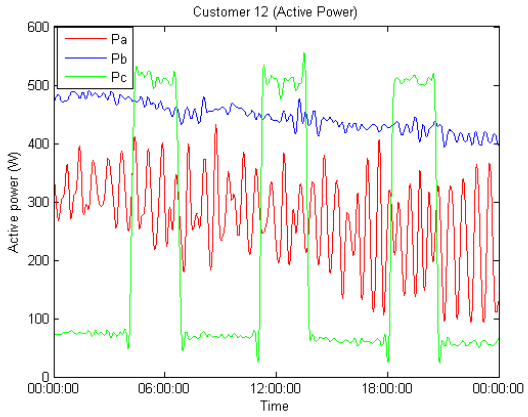


Figure 8. An example of load asymmetry.

V. DISCUSSION

The analyses on the state estimation accuracy were done based on the RTDS simulations although the proposed LV network monitoring concept was tested also in a real distribution network. In a real distribution network, the true state of the network is unknown and it is impossible to compare the estimated and the true values with the same accuracy as in RTDS simulations. There may also be other problems, such as inaccurate measurements and uncertainty about the network configuration. This is why simulations, as the ones done in this study, are a valuable help in state estimation research. The RTDS simulations were also vital for testing measurement devices, communication architecture and SS I-Dev software before the actual field test.

The real life tests were made in Italy in a LV network that contained single-phase customers and a significant amount of photovoltaic (PV) generation. The single-phase customers and PV generation were taken into account by making single-phase daily load profiles for customers with and without PV panels using smart meter measurements from the previous seven days. The short averaging window allowed the load profiles to follow the constantly changing PV production. PV production can change quickly as the solar irradiation increases during the spring and decreases during the autumn. Fig. 9 shows how the load profile for a single-phase customer with PV production changed at the beginning of February. The effect of increasing solar irradiation can be seen clearly even the study period is shorter than two weeks.

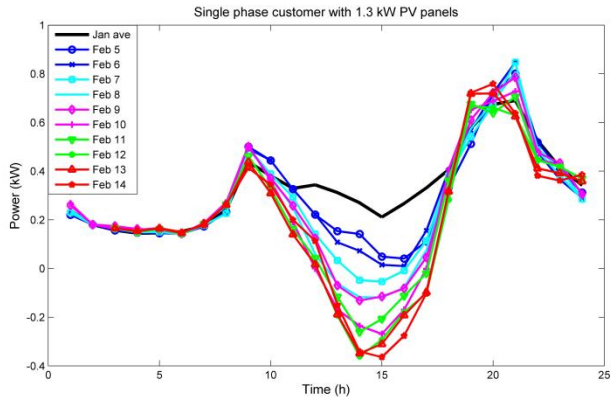


Figure 9. Load profile evolution during the beginnig of February.

In this case, it was easy to calculate the customer class load profile for a customer with PV production since all the customers belonged to the same customer class and had similar PV systems. If there had been different types of customers and PV systems of different sizes, the modelling task would have been more challenging. It would not be viable to model all possible load and PV production combinations, but we would have to model the load and PV production separately. For that, we would need a method for separating load and PV production from their sum measurement. If the load and PV production were modelled separately, it would be easy to scale the PV model according to the nominal PV system size and make short term PV production forecasts based on the solar irradiation forecasts.

VI. CONCLUSIONS

Load profiles, smart meter measurements and secondary substation measurements can be combined in a distribution network state estimator. The RTDS simulations described in this paper show that this improves the accuracy of low voltage network monitoring. Simulations also show that the state estimation accuracy increases if the meter reading frequency is increased and measurement averaging times are kept short. The decentralized distribution network management system proposed in INTEGRIS project enables fast and frequent communication between the state estimator and metering devices. When the state estimator provides accurate and almost real-time information on the low voltage network state, the automation level in the low voltage network can be increased. For example, unwanted loading situations could be avoided with the help of automatic congestion management, which would increase the lifetime and utilization rate of the low voltage networks.

REFERENCES

[1] M. Pikkarainen, A. Löf, S. Lu, T. Pöhö and S. Repo, ”Power quality monitoring use case in real low voltage network”, to be presented at the *4th European Innovative Smart Grid Technologies Conference.* Copenhagen, Denmark, 2013.

[2] S. Repo, D. Della Giustina, G. Ravera, L. Cremaschini, S. Zanini, J. M. Selga and P. Järventausta, ”Use case analysis of real-time low voltage network management,” presented at the *2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies*. Manchester, UK, 2011.

[3] A. Abur and A.G. Expósito, *Power System State Estimation: Theory and Implementation*. New York: Marcel Dekker, Inc. 2004.

[4] A. Abdel-Majeed and M. Braun, “Low voltage system state estimation using smart meters,” presented at the *47th Universities Power Engineering Conference*. London, UK, 2012.

[5] A. Abdel-Majeed, S. Tenbohlen, D. Schöllhorn and M. Braun, “Meter placement for low voltage system state estimation with distributed generation,” presented at the *22nd International Conference on Electricity Distribution*. Stockholm, Sweden, 2013.

[6] H. Wang, & N.N. Schulz, “A revised branch current-based distribution system state estimation algorithm and meter placement impact,” *IEEE Transactions on Power Systems*, Vol. 19, No. 1, February 2004. pp. 207-213.

[7] A. Mutanen, S. Repo and P. Järventausta, ”AMR in distribution network state estimation,” presented at the *8th Nordic Electricity Distribution and Asset Management Conference.* Bergen, Norway, 2008.

[8] A. Mutanen, A. Koto, A. Kulmala and P. Järventausta, ”Development and testing of a branch current based distribution system state estimator,” presented at the *46th International Universities' Power Engineering Conference*. Soest, Germany, 2011.

[9] T. Niknam and B.B. Firouzi, “A practical algorithm for distribution state estimation including renewable energy sources,” *Renewable Energy*, Vol. 34, Issue 11, November 2009. pp. 2309-2316.

[10] V. Thornley, N. Jenkins and S. White, “State estimation applied to active distribution networks with minimal measurements,” presented at the *15th Power System Computation Conference*. Liege, Belgium, 2005.

Publication 4

A. Mutanen, S. Repo, and P. Järventausta, "Customer classification and load profiling based on AMR measurements," presented at the 21st International Conference and Exhibition on Electricity Distribution (CIRED). Frankfurt, Germany, June 6–9, 2011.

Available at: http://www.cired.net/publications/cired2011/part1/papers/CIRED2011_0277_final.pdf

CUSTOMER CLASSIFICATION AND LOAD PROFILING BASED ON AMR MEASUREMENTS

Antti MUTANEN
Tampere University of Technology
Finland
antti.mutanen@tut.fi

Sami REPO
Tampere University of Technology
Finland
sami.repo@tut.fi

Pertti JÄRVENTAUSTA
Tampere University of Technology
Finland
pertti.jarventausta@tut.fi

ABSTRACT

Customer class load profiles are widely used in distribution network analysis. They are used, for example, in distribution network load flow calculation, state estimation, planning calculation and tariff planning. Previously, load profiling required expensive and time-consuming load research projects, but now automatic meter reading is providing huge amounts of information on electricity consumption. This paper presents different possibilities for utilizing AMR data on customer classification and load profiling. The customer classification and load profiling can be made separately or they can be combined by using clustering algorithms. Individual load profiles can also be formulated from the AMR measurements.

INTRODUCTION

Automatic meter reading (AMR) is becoming common in many European countries. In Finland, for example, distribution network operators (DNOs) are required to install AMR meters to at least 80 % of their consumption sites in their distribution networks by the end of 2013. Many DNOs plan to install AMR meters to all customers. AMR provides DNOs with accurate and up-to-date electricity consumption data. In addition to other functions, this data can be used to update load profiles and classify customers. The availability of AMR data also enables new and more accurate methods of modeling distribution network loads. Accurate load profiles are needed in daily used distribution network calculation, for example in load flow calculation, state estimation, planning calculation and tariff planning.

Distribution network customers are commonly classified to predefined customer classes, and the load of each customer is then estimated with customer class specific hourly load profiles. Currently, this method involves several error sources.

1) Sampling error. Parameters in the existing customer class load profiles can be based on measurements, which are misclassified or comprise an insufficient number of measurement points.
2) Geographical generalization. Load profiles are typically defined in national load research projects. Some of the accuracy is lost due to geographical generalization and within-country differences in electricity consumption are left unmodeled.
3) Profile drift. Electricity consumption is constantly changing but the load profiles are rarely updated.
4) Customer classification. DNOs have limited information on the type of the customers. The type of the customer is usually determined through a questionnaire when the electricity connection is contracted. However, the customer type may later change for instance because of a change in the heating solution.
5) Outliers. Some customers may have such an exceptional behaviour that they do not fit in any of the predefined customer class load profiles.

The above mentioned problems could be solved with the help of AMR measurements. The customer classification and load profiling could be done according to actual consumption data. Since AMR data is collected continuously, the classification and load profiles would remain up-to-date at all times. The classification and accuracy of the load profiles could be checked automatically for instance once a year. The load profiles could also be calculated separately for each DNO or region, thus avoiding the errors caused by geographical generalization. Outliers could be detected and individual load profiles could be formed for the outliers. Individual load profiles could also be calculated for some of the largest customers to improve the load estimation accuracy.

In this paper, we use real AMR data to update customer class load profiles and reclassify customers. Different classification methods, from simple reclassification to existing customer classes to K-means and ISODATA clustering (Iterative Self-Organising Data Analysis Technique), are tested. The results are compared with the original customer classification and load profiles. This paper shows that updated DNO specific customer class load profiles have a big effect on the accuracy of the load estimates. Furthermore, a method for forming individual load profiles for outliers or large customers is presented.

BACKGROUND AMR DATA

Two different measurement sets from two Finnish distribution companies are used to study the different possibilities for utilizing AMR data on customer classification and load profiling. The first measurement set contains AMR measurements from 127 residential customers. These measurements cover the years 2006–2007. The second measurement set contains interval

measurements from 660 customers. All of these interval metered customers have annual energy consumption larger than 100 MWh/year. The measurement period for the interval metered customers was from 18 August 2008 to 31 December 2009. Both measurement sets have one hour measurement interval.

Here, the first year of measurement data is used for customer classification and load profiling and the rest of the data is used for the verification of the results. The residential measurements are used for load profile updating, reclassification, clustering and individual load profiling. The interval measurements are used for studying clustering and individual load profiles.

One year of measurement data is the minimum requirement for customer classification and load profiling. Better results are achieved if more data is available. However, if a lot of data is available, the possible changes in electricity consumption should be taken into account by weighting the most resent years or detecting change points.

CUSTOMER CLASSIFICATION

Distribution system loads are commonly estimated with customer class load profiles. Each customer is linked to one of the predefined customer classes, and the load of each customer is then estimated with the customer class specific hourly load profile [1]. This method assumes that the distribution system operator knows which customer belongs to which customer class. In practice, classification errors are common.

AMR measurements can be used to improve the customer classification accuracy. Every customer with AMR can be classified according to its actual consumption by comparing the measured electricity consumption with the customer class load profiles or other customers. The customer classification can be made in many ways. The customers can be simply reclassified to the nearest existing customer class load profiles or new customer classes can be formed by grouping customers with similar behaviour. A simple reclassification procedure is defined and studied. Some test results are described in the following.

Case 1: Customer reclassification

In customer reclassification, AMR measurements are used to determine which existing customer class load profile is closest to customer's true load pattern. Then the customer is reassigned to this customer class. The 127 residential customers studied in this paper are reclassified according to AMR measurements from the year 2006. Euclidian distance between the measurement and customer class load profile is used as a distance measure.

The studied customers were originally divided into six customer classes. They belonged to a network company which uses 38 customer classes. Table 1 shows how the customers were divided into the existing customer classes before and after customer reclassification.

After customer reclassification, the studied customers were scattered to 14 different customer classes. The accuracy of the customer classification was measured by using the customer class load profiles to make next day electricity consumption forecasts for the year 2007. Square sum of the forecast error was calculated for both original and new customer classification. Compared to the initial situation, the customer reclassification reduced the square sum of forecast errors by 7 %. The results can also be seen in Figure 1.

Here, as in the following cases, the outdoor temperature was taken into account with four season specific temperature dependency factors. The load profiles model the load in long-term average temperature. When making the next day load forecasts, the load was corrected according to the next day average temperature (forecast, in real applications).

LOAD PROFILE UPDATING

Previously, load profiling required expensive and time-consuming load research projects and therefore the load profiles were rarely updated. Old load profiles and the constant change in electricity consumption habits have caused significant profile drift to the customer class load profiles. During the last decade the use of entertainment electronics has increased, heat pumps and air conditioners have become more common and lighting efficiency has increased, just to name a few changes.

AMR measurements could be used to update customer class load profiles. This would have several benefits. Regularly, for instance once a year, done load profile update would keep the load profiles up-to-date at all times. This would ensure that the load profiles keep up with the changing electricity consumption habits. Also, errors that are associated with sampling and geographical generalization would decrease. The sampling errors decrease when measurements from all or almost all customers are used in the load profile calculation. The geographical generalization could be avoided by calculating the load profiles separately for each distribution network area or region.

Case 2: Load profile update

Load profile updating was studied here as an alternative to the customer reclassification. Six updated customer class load profiles were calculated for the 127 residential customers previously studied in case 1.

Table 1. *Customer classification before and after customer reclassification.*

Customers per customer class	1	2	3	4	5	6	7	8	10	11	15	26	28	30	31	38
Original classification	30	41	43	3	-	-	7	-	-	-	-	-	3	-	-	-
Updated classification	3	37	14	12	4	3	-	1	22	8	4	3	-	14	1	1

The load profile update was done with measurement data from the year 2006 using the original customer classification. As shown in Figure 1, the load profile update provided a 30 % reduction to the overall square sum of the forecast error.
The load profile update had a bigger effect on the load forecasting accuracy than the customer reclassification. The load profile update and the customer reclassification should of course be combined to achieve the best result. However, if the load profile update is done after the customer reclassification, the updated customer class load profile is no longer the nearest load profile for all customers. The customer class reassignment and load profile update should be done again and again until none of the customers change customer class during the reclassification process. Basically, this is a clustering problem. Clustering is studied in the next chapter.

CLUSTERING

Clustering is an efficient technique for finding customers with similar behaviour. In literature, several different clustering methods have been applied to electricity customer classification [2], [3]. In this study, K-means and ISODATA clustering algorithms are used to solve the customer classification problem.
The clustering is done based on the AMR measurements, but since the hourly measurement data has a very high dimensionality, some kind of a dimension reduction is needed to speed up the computation and to get feasible results. There are many techniques for dimension reduction, for example principal component analysis, Sammon map and curvilinear component analysis [2]. Here, a pattern vector approach is used. The whole year's electricity consumption is described in a pattern vector containing average weekly loads for each calendar month. The pattern vector describes daily, weekly and monthly load variations on an hourly basis. In addition to 2016 hourly load values, the pattern vectors also include four customer specific temperature dependency parameters.
More information on the pattern vector formation and used ISODATA algorithm can be found in reference [4].

Case 3: Clustering residential customers

Both K-means and ISODATA clustering algorithms were used to cluster the studied 127 residential customers into six customer groups. After clustering, new updated customer class load profiles were calculated for each customer class. Square sums of the forecast errors were calculated as before and the results were compared. K-means and ISODATA algorithms provided very similar customer classification accuracy. Figure 1 shows that both of these clustering methods reduced the square sum of errors by 36 % compared to the initial situation.
Classification accuracy was similar, but the K-means clustering was found out to be a lot simpler to execute than the ISODATA clustering.

Case 4: Clustering non-residential customers

660 interval metered customers were used to demonstrate the clustering of non-residential customers. Before clustering the outliers were filtered from the data set. There is no point in trying to cluster customers whose electricity usage differs significantly from the other customers. Instead, individual load profiles can be formed for the outliers. Two stage statistical filtering was applied. The filtering was done based on monthly energy consumptions and Euclidian distances between pattern vectors. 92 customers were classified as outliers. More information on the used outlier filtering procedure can be found in reference [4]. No outlier filtering was done in Case 3 to keep the results comparable with Case 2.
The pattern vectors used in clustering were formed from measurements between 18 August 2008 and 17 July 2009.

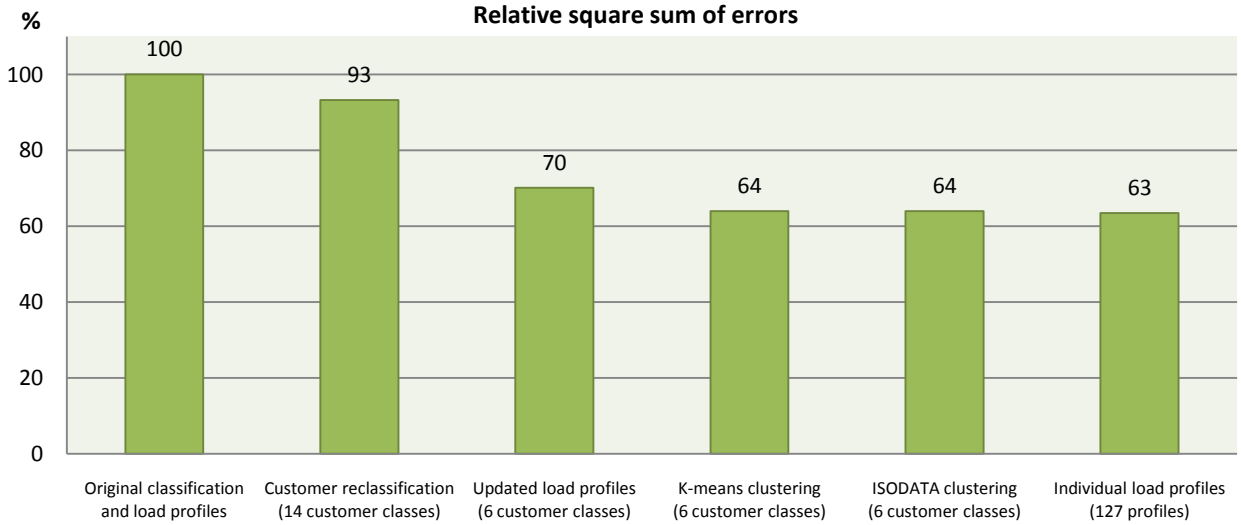


Figure 1. *Square sum of the forecasting errors for 127 residential customers, in relation to the original situation.*

The 568 customers who passed the outlier filtering were clustered into 30 clusters and customer class load profiles were calculated for each cluster. Next day electricity consumption forecasts were made for the time period 18 August to 31 December 2009. The forecasting accuracy is later compared with the forecasting accuracy of individual load profiles in Figure 2.

INDIVIDUAL LOAD PROFILES

Now that AMR measurements are commonly available, many DNOs are thinking of replacing the customer class load profiles with previous year's AMR measurements. In fact, the DNO that supplied the interval measurements for this study is already modeling interval metered customers that way. Previous year's measurements without temperature or special day correction are used as reference models in Figure 2.

When using measurements to model individual loads, we should take into account the facts that even consecutive years are not identical and individual loads are highly stochastic in nature. If the measurements are used for making load forecasts, the random variations in the weather and customers' hourly electricity consumption should be taken into account. The outdoor temperature can be taken into account with customer specific temperature dependency factors. In short-term forecasting the temperature forecasts can be used to adjust the load level and average temperatures can be used in long-term forecasts.

In current (Finnish) customer class load profiles the profiling errors and stochastic variations in hourly loads are described with standard deviation. The same approach should be applied also to the AMR measurement based individual load profiles.

In this study, individual load profiles are formed from measurements by calculating representative type weeks for each month. This method smoothes out the stochastic variations on hourly loads and enables the calculation of standard deviations. In type week, each day of the week is modeled separately. Holidays are modeled as Sundays.

Case 5: Residential customers

Individual load profiles were formed for the 127 residential customers previously studied in Cases 1–3. As depicted in Figure 1, the forecasting accuracy of individual load profiles was only marginally better than the accuracy achieved with clustering and customer class load profiles.

Case 6: Non-residential customers

With non-residential customers, the individual load profiles provided better results. The square sum of the forecasting errors decreased about 17 % compared to the clustering methods in Case 4. The type week based individual load profiles were 21 % more accurate than the load models based directly on the previous year's measurements. The results are also shown in Figure 2.

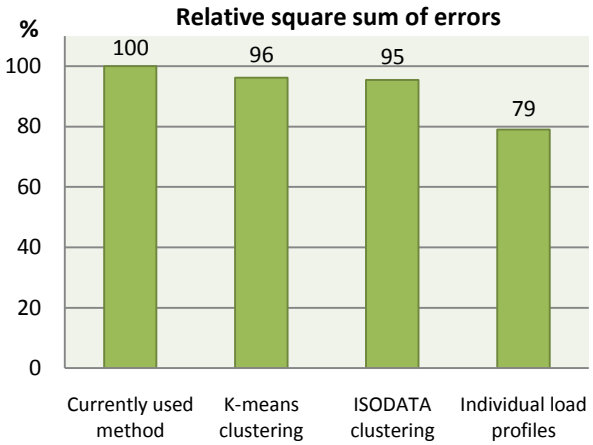


Figure 2. *Relative square sum of the forecasting errors for 568 interval metered customers.*

CONCLUSIONS

This paper compared different methods for utilizing AMR data on customer classification and load profiling. The simple customer reclassification to existing customer classes provided little improvement to the load profiling accuracy. Calculating updated DNO specific customer class load profiles was a much more efficient method to improve the load profiling accuracy. However, even better results were achieved by combining the customer reclassification and load profile updating with clustering methods.

The use of individual load profiles was also studied. When studying small residential customers, the individual load profiling improved the load profiling accuracy only marginally compared with the clustering methods. Only in the case of large non-residential customers, the accuracy improvement was large enough to make individual load profiling a viable option.

REFERENCES

[1] A. Seppälä, 1996, "Load research and load estimation in electricity distribution", Ph.D. dissertation, Helsinki University of Technology.

[2] G. Chicco, R. Napoli, & F. Piglione, 2006, "Comparison Among Clustering Techniques for Electricity Customer Classification", *IEEE Trans. Power Systems*, vol. 21, 933–940.

[3] G.J. Tsekouras, N.D. Hatziargyriou, & E.N. Dialynas, 2007, "Two-Stage Pattern Recognition of load Curves for Classification of Electricity Customers", *IEEE Trans. Power Systems*, vol. 22, 1120–1128.

[4] A. Mutanen, M. Ruska, R. Repo, P. Järventausta, 2010, "Customer Classification and Load Profiling Method for Distribution Systems", Manuscript submitted for publication.

Publication 5

A. Mutanen, M. Ruska, S. Repo, and P. Järventausta, "Customer classification and load profiling method for distribution systems," IEEE Transactions on Power Delivery, vol. 26, no. 3, July 2011.

Available at: https://doi.org/10.1109/TPWRD.2011.2142198

Customer Classification and Load Profiling Method for Distribution Systems

Antti Mutanen, Maija Ruska, Sami Repo, and Pertti Järventausta

***Abstract*—In Finland, customer class load profiles are used extensively in distribution network calculation. State estimation systems, for example, use the load profiles to estimate the state of the network. Load profiles are also needed to predict future loads in distribution network planning. In general, customer class load profiles are obtained through sampling in load research projects. Currently in Finland, customer classification is based on the uncertain customer information found in the customer information system. Customer information, such as customer type, heating solution and tariff, is used to connect the customers with corresponding customer class load profiles. Now that the automatic meter reading systems are becoming more common, customer classification and load profiling could be done according to actual consumption data. This paper proposes the use of the ISODATA algorithm for customer classification. The proposed customer classification and load profiling method also includes temperature dependency correction and outlier filtering. The method is demonstrated in this paper by studying a set of 660 hourly metered customers.**

***Index Terms*—Clustering, ISODATA, K-means, load profiles, load research.**

I. Introduction

In Finland, distribution system loads are commonly estimated with load profiles. Each customer is linked to one of the predefined customer classes, and the load of each customer is then estimated with customer class-specific hourly load profiles. The method involves several error sources and presents significant uncertainties in load estimation. Classification errors are common, because customer classification is based on uncertain customer information. The type of the customer is usually determined through a questionnaire when the electricity connection is contracted. Once the customer type has been determined, it is hardly ever updated. In reality, the customer type may change, for instance, because of a change in the heating solution or an addition of new devices, such as air conditioning. It is a difficult and sometimes impossible task for the system operator to detect the change in customer type only based on billing information.

Moreover, the parameters in existing customer class load profiles can be based on measurements, which are old, misclassified or comprise an insufficient number of measurement points. This is also a significant error source.

Even if the customer information needed in the classification is correct, some of the customers can simply have such an irregular behaviour pattern that they do not fit in any of the predefined customer class load profiles. The predefined customer class load profiles also include some inaccuracy due to geographical generalization. The most widespread customer class load profiles are created to model the average Finnish electricity consumption. They do not take into account the regional differences in electricity consumption, which originate from different climate conditions and socioeconomic factors.

Automatic meter reading (AMR) is becoming common in many European countries. AMR provides distribution system operators (DSOs) with accurate and up-to-date electricity consumption data. These data can be used to classify and model distribution network loads. The amount of load data will be enormous when all or almost all of the customers have hourly metering. Since one DSO can have several hundreds of thousands of customers, some kind of automatic data analysis and clustering method should be used.

This paper proposes a pattern recognition method for customer data classification. The method classifies customers into clusters, for which load profiles can be calculated. These profiles are then used to model customer loads in the distribution system. The method involves temperature dependency correction and outlier filtering.

Different types of clustering techniques have been proposed in the literature for customer classification and load profiling. For example, classical clustering and statistical techniques [1]–[6]; data mining [7], [8]; self-organizing maps [1], [2], [4], [9]; neural networks [10], [11]; and fuzzy logic [4], [5], [10]–[12] have all been applied before.

In the previous studies, the customer classification has typically been made according to daily load profiles or load shape factors. Here, the classification is made according to pattern vectors which include daily, weekly, monthly, and seasonal load variations. Also the motive for customer classification is different. Previously, classification has been studied for the purpose of tariff formulation or marketing

Manuscript received August 16, 2010.

A. Mutanen, S. Repo, and P. Järventausta are with the Department of Electrical Energy Engineering, Tampere University of Technology, P.O. Box 692, FI-33101 Tampere, Finland, (e-mail: antti.mutanen@tut.fi; sami.repo@tut.fi; pertti.jarventausta@tut.fi).

M. Ruska is with the VTT Technical Research Centre of Finland, P.O. Box 1000, FI-02044 Espoo, Finland, (e-mail: Maija.Ruska@vtt.fi).

strategy planning. Here the main incentive has been the need for more accurate network calculation: distribution network state-estimation [13] and network planning calculation.

The current trend in electricity distribution is to maximize the quality of supply and utilization degree of the existing networks with the help of active network management. Advanced distribution automation functions, such as coordinated voltage and reactive power control, automatic feeder reconfiguration and load control, require accurate voltage and power flow estimates. Load model accuracy has a big effect on the distribution network state estimation accuracy [13].

The presented classification method was developed at the VTT Technical Research Centre of Finland. VTT has also developed an application utilizing the presented classification and load profiling method. The LoadModellerPRO program composes load profiles automatically from AMR data and is used by several Finnish distribution system operators. In this paper, the classification and load profiling method is transferred to the MATLAB environment and its classification accuracy is reviewed by comparing it to alternative classification methods.

The presented classification method is universal and can be applied wherever there is sufficient AMR data available. Only the load profiling method needs to be modified to suit local needs and practices. A Finnish case study is presented here. The Finnish distribution system environment provides an excellent platform for the presented method. The hourly load profiles have been in use for a long time and the AMR installations are increasing rapidly. Finnish DSOs are required to equip at least 80 % of the customers with AMR by the end of the year 2013. Section II describes the current Finnish customer classification and load modeling practices. The developed classification method is presented in Section III. Section IV presents some results and Section V discusses the use of the presented classification method. Finally, conclusions are given in Section VI.

II. Load Modeling Method

Finnish load research tradition dates back to the 1980s, when DSOs started to cooperate in load research. The structure of the load model was developed more than 20 years ago. A short description of the Finnish load modeling method is given in Sections II-A, II-B and II-C. In-depth information can be found in [3].

DSOs have customer information systems (CISs), which store all the available information of each customer's electrical connection, type and electricity consumption. The customer data usually include:

- Electricity connection information: customer location, supply voltage, fuse size, number of phases;
- Customer class: residential, agriculture, public, service, industry (NACE code or some other similar code indicating the line of business);
- Consumption: annual electricity consumption, high and low tariff electricity consumption (if dual time tariff);
- Additional information: heating system (in the case of electric heating: type of electric heating), type of domestic hot water heating system, existence of electric sauna stove.

Traditionally, distribution system estimation uses customer class load profiles for load modeling. Using the information from CIS, each individual customer is linked to one predefined customer class load profile. Finnish DSOs usually use approximately 20–50 customer classes. In addition, some of the largest customers are often modelled with their own models. The customers are also linked to the geographic network model in the network information system (NIS). This enables network calculations using the load profiles.

A. Model Structure

The load model used nowadays by most Finnish DSOs' software applications represents the expectation value E[$P(t)$] and standard deviation $s_P(t)$ for the customer's hourly load as a linear function of the annual energy consumption W_a. The load model can be represented either as topography or as an index series. In topography, the expectation value and standard deviation for hourly load are given for every hour of the year. The expectation value L_{topo} and standard deviation s_{topo} are usually given for a base energy consumption of 10 MWh/year (W_{base}).

In index series, the load parameters are given in a relative form. The index $Q(t)$ models seasonal variation with 26 two-week indices. The index $q(t)$ models hourly variation for three different day types (working day, Saturday, and Sunday). Each two-week period is modelled separately in index $q(t)$, which thereby consists of 26*3*24=1872 indices. Overall, the load expectation values for the whole year are modelled with 1872+26=1898 parameters. The hourly standard deviations for the three day types are given as a percentage of the average load in the index $s_{\%}(t)$.

Formulas for calculating the hourly load parameters (expectation values and standard deviations) with topographies (1) and index series (2) are given below.

$$\begin{cases} \mathrm{E}[P(t)] = L_{topo}(t) \cdot W_a / W_{base} \\ s_P(t) = s_{topo}(t) \cdot W_a / W_{base} \end{cases} \quad (1)$$

$$\begin{cases} \mathrm{E}[P(t)] = \dfrac{W_a}{8760} \cdot \dfrac{Q(t)}{100} \cdot \dfrac{q(t)}{100} \\ s_P(t) = \mathrm{E}[P(t)] \cdot \dfrac{s_{\%}(t)}{100} \end{cases} \quad (2)$$

Topographies take special holidays into account, but in the index series, public holidays and eves are modelled as Sundays and Saturdays, respectively. Both in topographies and in index series, the reactive power is calculated using one customer class-specific power factor for every hour of the year. In some distribution companies, the reactive power is modeled like the active power with topographies or index series.

B. Utilization of Load Models

In Finland, loads are modeled down to the individual customer level. Every customer is connected into the network data even at the low-voltage (400 V) level. In distribution network calculation, the customer-level loads are aggregated into higher level loads according to probability theory. For simplicity, loads are assumed normally distributed and independent. In that case, the aggregated load expectation values $\mathrm{E}[P_{ag}(t)]$ and standard deviations $s_{ag}(t)$ for n customers can be calculated with (3) and (4) [3].

$$\mathrm{E}\left[P_{ag}(t)\right] = \mathrm{E}[P_1(t)] + \mathrm{E}[P_2(t)] + \ldots + \mathrm{E}[P_n(t)] \tag{3}$$

$$s_{ag}(t) = \sqrt{s_1(t)^2 + s_2(t)^2 + \ldots + s_n(t)^2} \tag{4}$$

The stochastic nature of the loads is taken into account when calculating peak loads. Load values with different excess probability levels are used in distribution network calculation. The load $P_p(t)$ having an excess probability of p % can be calculated with (5).

$$P_p(t) = \mathrm{E}[P(t)] + z_p \cdot s_P(t), \tag{5}$$

where z_p is the standard normal deviate corresponding to excess probability p. The load values with excess probability around 10 % are relevant for voltage-drop calculation, while smaller probabilities are used when studying loading limits. The load expectation values are used when calculating losses. [14]

C. Weather Dependency

The influence of weather on electricity demand is a widely studied phenomenon [15]. Outdoor temperature is usually the single most important factor, but also wind and cloudiness affect electricity demand. In distribution network calculation, a simple weather dependency model is adopted, and only the outdoor temperature dependency is taken into account. In Finland, different electric heating options are widespread, and this, combined with large temperature variations, renders the modeling of the temperature dependency essential in the statistical analysis of customer loads.

As individual loads are metered in different time and location, the effect of temperature variation on a load should be screened out of the data before customer classification. In Finland, a simple and robust model for temperature dependency has been adopted. The temperature-dependent part of the load is modelled as

$$\Delta P(t) = \alpha \cdot (T_{ave} - \mathrm{E}[T(t)]) \cdot \mathrm{E}[P(t)], \tag{6}$$

where $\Delta P(t)$ is the outdoor temperature dependent part of the load P at time t;
T_{ave} is the average temperature of the previous day;
$\mathrm{E}[T(t)]$ is the expectation value of the outdoor temperature at time t (long-term daily average temperature);
α is the seasonal temperature dependency parameter [%/°C];
$\mathrm{E}[P(t)]$ is the expectation value of the load at time t.

In this paper, the parameter α is calculated with linear regression analysis for every four seasons separately. Daily energy consumptions and daily average temperatures are used in the analysis. The effects of daily and monthly fluctuations in electricity demand are eliminated by choosing the regressand and regressor as follows:

- Regressand: the percent error between the daily energy consumption and the average daily energy consumption on a similar day (same weekday and month).
- Regressor: difference between the daily average temperature and the average temperature on a similar day. A one day delay was added to the daily average temperatures to account for the delay in temperature dependency [15].

III. Clustering Method

As the classes and the number of classes are not known beforehand, an unsupervised classification method should be used. In this paper, iterative self-organising data analysis technique (ISODATA) algorithm is used. The algorithm allows the number of clusters to be automatically adjusted, if needed.

A. Pattern Vectors

Before the clustering algorithm is applied, each customer's metered load is transformed to a pattern vector. The vector consists of four temperature dependency parameters and 2016 hourly load values. The seasonal temperature dependency parameters are calculated individually for each customer. The achieved parameters are used to normalize the metered load to long-term average temperature. The load values contain weekly average loads calculated for each calendar month.

The annual energies of different customers can vary greatly. The load values in pattern vectors are normalized by dividing each load element by the vector's average load.

B. Outliers

At this stage, outliers are distinguished from other data. Outliers can be failed measurements or customers who use electricity in a very different way from average customers. Two main types of the outliers are:

1) Customers whose electricity use varies significantly during some months. These are detected by comparing each individual customer's monthly energies to the all customers' average monthly energy. If a customer's monthly energy differs from the average more than is probable with probability p from the normal distribution, the customer is an outlier. Probabilities between 80% and 99.99% can be applied for this calculation.

2) Customers whose intra-day load variation is very high compared to other customers. These customers are filtered out with the help of Euclidean distance measure. The calculation of the Euclidean distance of a pattern vector is described later in Section III-C. If individual customer's Euclidean distance from all customers' average vector is larger than what is probable with probability p from the normal distribution, the customer is classified as an outlier. Probabilities between 80% and 99.99% can be applied for this calculation.

C. Clustering Algorithm

Euclidean distance (7) is chosen for the similarity measure used in the clustering algorithm. The Euclidean distance between two n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \qquad (7)$$

The first four parameters in the pattern vector are temperature dependency parameters. These parameters are weighted in the analysis. Suitable weights are found experimentally. The weight of the temperature dependency parameters is defined as 5 % and the weight of the actual load measurements is defined as 95 %.

The pattern vectors are clustered using the ISODATA algorithm. The method includes heuristic provisions for splitting an existing cluster into two and for merging two existing clusters into a single cluster. The method is unsupervised—the user need not to know the exact number of classes before clustering is completed.

The main procedure of the algorithm is (see for example [16] or [17]):

1) Cluster the existing data into c classes but eliminate any data and classes with fewer than T members and decrease c accordingly (Procedure 1). Exit when classification of the samples has not changed.
2) If $c \le c_d/2$ or $c < 2c_d$ and iteration odd, then
   a) Split any clusters whose samples form sufficiently disjoint groups and increase c accordingly (Procedure 2).
   b) If any clusters have been split, go to step 1.
3) Merge any pair of clusters whose samples are sufficiently close and/or overlapping and decrease c accordingly (Procedure 3).
4) Go to step 1.

Here, c is the number of clusters, c_d is the desired number of clusters, and T is the minimum number of samples in a cluster.

Procedure 1 is a variant of the K-means procedure [18]. A flowchart is shown in Fig. 1.

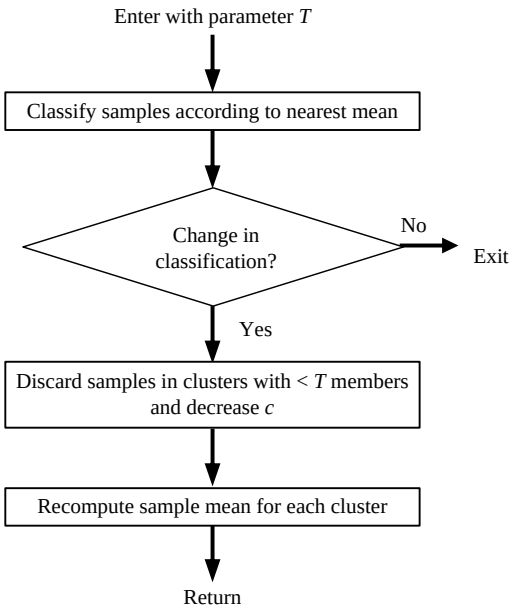


Fig. 1. A flowchart for Procedure 1.

Procedure 2 for splitting is somewhat heuristic. The flowchart is given later in Fig. 2. ISODATA replaces the original cluster centre with two centres displaced slightly in opposite directions along the axis of largest variance.

The splitting procedure is always performed when the number of clusters is smaller than half the desired number of clusters. Splitting is not performed if the number of clusters is at least twice the desired number of clusters. When the number of clusters is within range ($c_d/2$, $2c_d$), splitting is performed every second round. The desired number of clusters is given by the user and it defines the approximate number of clusters wanted. The final number of clusters also depends on the other user given parameters and the natural number of clusters in the data.

Two different measures d_k and S_k are used to evaluate the uniformity of the clusters. The quantity d_k is the average distance of samples from the mean of the kth cluster and the S_k is the sum of the largest squared distances from the mean along the coordinate axes. Note that here the latter uniformity measure differs from the original (presented in [16] or [17]). Originally, this uniformity was described with a value calculated from only one coordinate axis. In customer load data classification, the uniformity of clusters is better described with information from all coordinate axes.

$$d_k = \frac{1}{N_k} \sum_{\boldsymbol{x} \in \chi_k} d_E(\boldsymbol{x}, \boldsymbol{m}_k), \qquad k = 1,2,\ldots,c \qquad (8)$$

$$S_k = \frac{1}{N_k} \sum_{i=1}^{n} \max_j \left(x_i^{(j)} - m_{ki}\right)^2, \qquad k = 1,2,\ldots,c \qquad (9)$$

where N_k is the number of samples in cluster k,
χ_k is the set of vectors belonging to cluster k,

$\boldsymbol{m}_k$ is the average vector of cluster k,
$d_E(\boldsymbol{x}, \boldsymbol{m}_k)$ is the distance of vector $\boldsymbol{x}$ from cluster k's average vector,
n is the number of elements in pattern vector,
$x_i^{(j)}$ is the ith element of pattern vector $\boldsymbol{x}^{(j)}$ belonging to cluster k (j=1,2,…,N_k),
m_{ki} is the ith element of $\boldsymbol{m}_k$.

The overall average distance of samples d is defined by

$$d = \frac{1}{c}\sum_{k=1}^{c} N_k d_k . \qquad (10)$$

The cluster is split, if the sum of largest squared distances from the mean of the cluster k (S_k) is larger than the user defined threshold value S_s ($S_k > S_s$) and

$$[d_k > d \text{ and } N_k > 2(T+1)] \text{ or } c < \frac{c_d}{2} . \qquad (11)$$

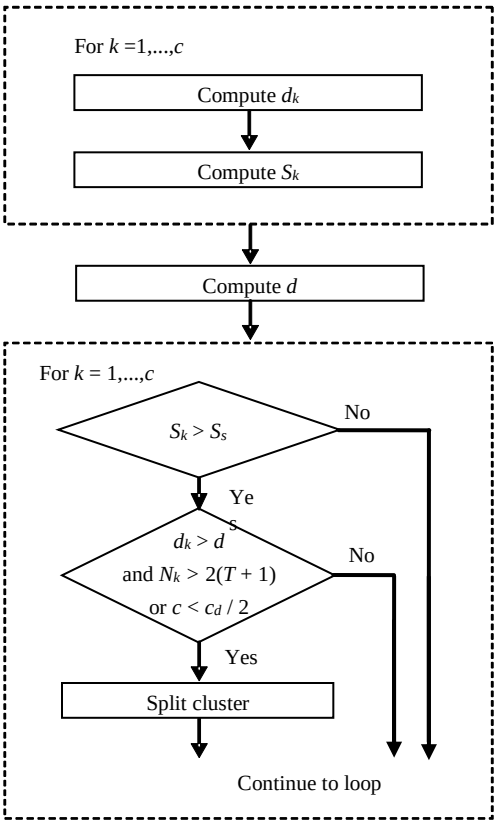


Fig. 2. A flowchart for Procedure 2 (splitting).

Procedure 3 for merging is performed only if splitting is not executed. Procedure 3 for merging is shown in Fig. 3. At first, all pairwise distances between cluster centers d_{ij} are calculated and compared to the threshold value D. Those pairs of clusters corresponding to distances that are less than the threshold value D are arranged in a list from the smallest distance to the largest. The clusters are then merged according to the list's order. Merging continues as long as the total number of merges does not exceed the maximum limit (input parameter M_{max}).

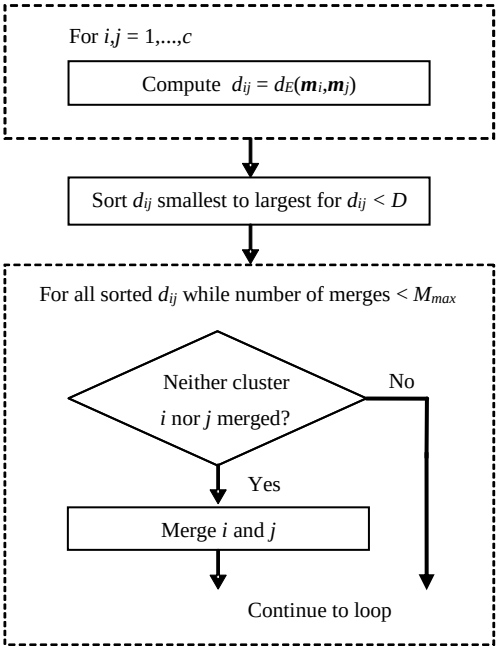


Fig. 3. A flowchart for procedure 3 (merging).

IV. Results

The algorithm shown before was written into a MATLAB program, and its performance is studied here using a set of measurements from 660 hourly measured customers. The measurements have been acquired from a distribution network company in Western Finland. The measurement period used in customer classification and load profiling is from August 18, 2008 to August 17, 2009. The available hourly electricity consumption data had 1-kWh/h measurement resolution. Therefore, only large customers with annual energy consumption larger than 100 MWh/year are studied. Hourly temperature measurements were also available for the studied network area.

A. Measurement Pre-Processing and Outlier Filtering

The measured electricity consumption data can contain errors due to faults in metering or communication. Also, data format changes can cause errors. Typically, these errors are seen as missing values or as errors in the order of magnitude.

In this study, the following pre-processing rules were applied: if the measurement contained a missing data interval longer than five hours or the number of the missing data intervals was larger than five the measurement was omitted from the data set. Missing parts of the data were estimated using linear interpolation. If the measured hourly value was clearly of wrong magnitude, the right order of magnitude was estimated by comparing it with the magnitude of the previous hourly value.

Next, the pre-processed measurements were normalized to

long-term (30 years) average temperature, and the individual pattern vectors were formed. Then, the measurements were grouped into six different main customer classes according to the customer class information found in CIS. The selected main customer classes were: residential customers (private apartments and housing corporations combined), agricultural customers, industrial customers, public administration, commercial customers, and other customers (combination of construction, traffic, lighting, and community management).

The outlier filtering was accomplished according to the method presented in Section III-B. A 99% probability level was used to detect abnormalities in monthly energy consumption and a 95% probability level was used to detect abnormal intra-day load variations. Examples of the filtered pattern vectors can be seen in Figs. 4 and 5. Note that even if a pattern vector gets filtered, it does not necessarily mean that the corresponding measurement is erroneous; the customer may simply have an extraordinary load pattern.

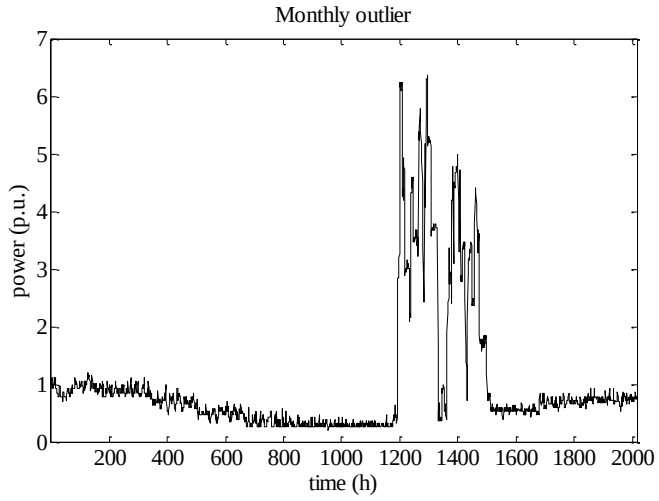


Fig. 4. Pattern vector for a customer with exceptionally large monthly energy consumption in August and September (only the load part of the pattern vector is shown).

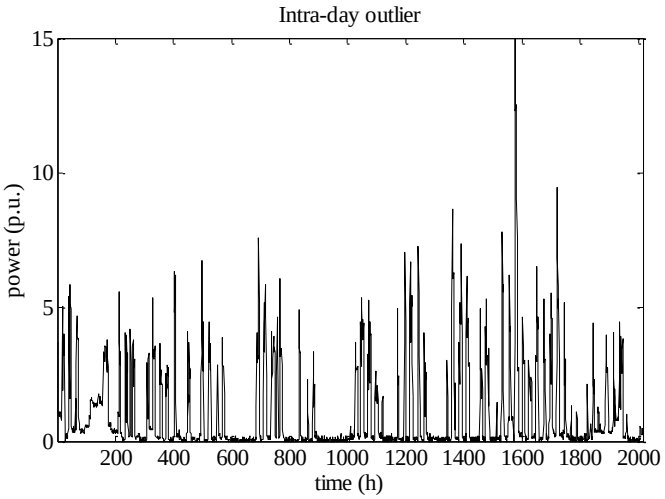


Fig. 5. Pattern vector for a customer with abnormal intra-day behaviour (only the load part of the pattern vector is shown).

The outlier filtering procedure classified 92 out of 660 pattern vectors as outliers.

B. Clustering

The clustering algorithm introduced in Section III-C was used to cluster the remaining 568 pattern vectors. The clustering procedure was carried out separately for each main customer class. Fig. 6 presents the clustering results for the public administration main customer class. For clarity, only the week corresponding consumption in January is presented. The cluster centres are marked with bold black lines and the individual pattern vectors are marked with gray lines. The following parameters were used when clustering public administration customers: c_d=4, T=1, S_s=30, D=11, and M_{max}=5.

The clustering algorithm divided the 127 pattern vectors in the public administration main customer class into five distinct clusters. The number of pattern vectors (n_k) in each cluster varied between 9 and 50. The public administration main customer class contained a total of 151 customers, 24 of them were classified as outliers in the previous step.

Once the classification of the customers is completed, the customer class load profiles can be calculated. The hourly load profiles can be calculated from the original temperature-normalized measurements. The load profiles can be expressed either as topographies or as index series.

Individual load profiles should be used for the outliers. We recommend that the individual load profiles are formed with the same principle as the pattern vectors. That is, the day-type-specific monthly averages are used as expectation values. The use of monthly averages helps to smooth out the effect of stochastic variation in the load expectation values. Also, the standard deviations can be calculated when each value is a mean of approximately four hourly values.

The standard deviation calculation is not really reliable if the sample only consists of four hourly values. However, if measurement data are available only from a period of one year, this is a simple way to produce a rough estimate for the standard deviation. After the standard deviations have been calculated, the individual load profiles can be expressed as topographies or index series. In topographies, the average load profile describing one week's consumption is simply duplicated to cover the whole month.

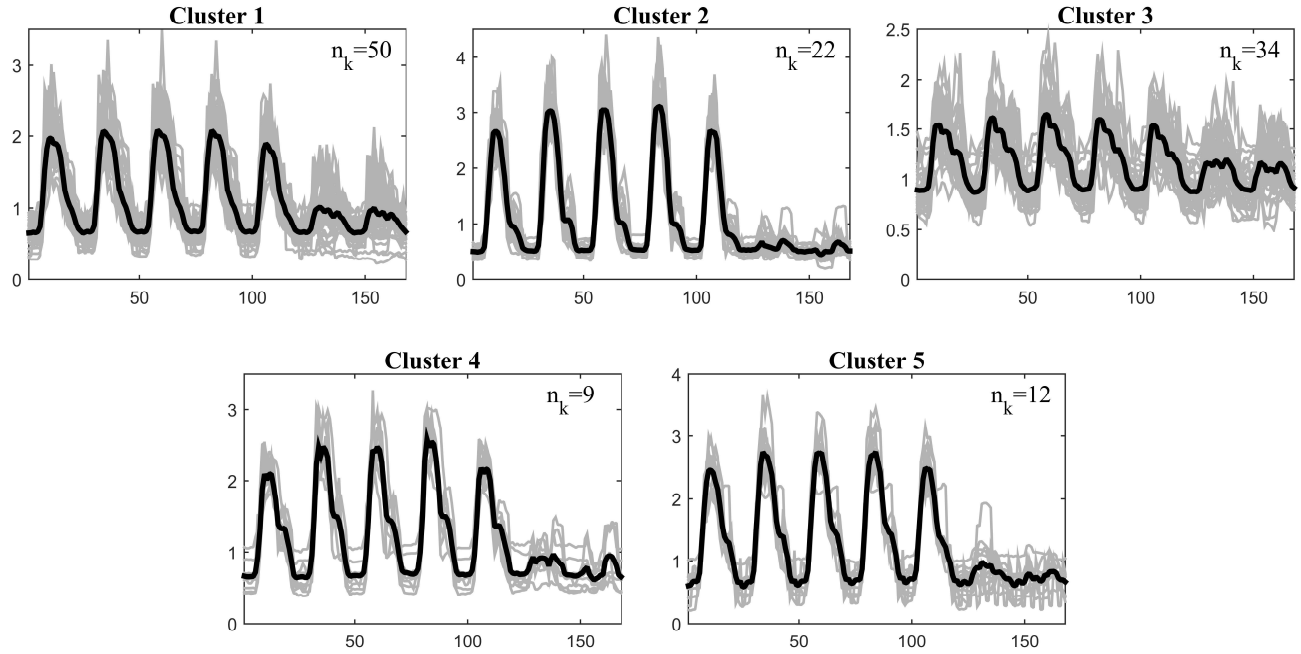


Fig. 6. Results of ISODATA clustering for the public administration main customer class. Horizontal axis: time (h); vertical axis: normalized load.

The accuracy of load profiles could be increased by increasing the number of customer classes. However, in practice a compromise between accuracy and number of customer classes has to be made. Here the desired number of clusters c_d was selected on the basis of the knee point criterion [1]. The knee point criterion helps to find the optimal number of clusters. Fig. 7 shows how the public administration load profile square sum of errors (SSE) between the cluster centres and the measurements depends on the number of the clusters. The knee point is roughly in four clusters. The SSE values in Fig. 7 are calculated similarly as in Section IV-C. For simplicity, the K-means clustering algorithm was used instead of ISODATA when searching for the knee points. In practice, the operator selects the desired number of clusters empirically.

The other user given parameters also affect the number of clusters. The thresholds for splitting and merging (S_s and D) define how many times the clusters are split and merged. Choosing the right threshold values requires advance information on the type of the customers or use of trial-and-error technique. High threshold values are chosen when clustering customers with high stochasticity and low thresholds are chosen when clustering customers with low stochasticity. Also the number of customers affects the threshold values. Table I shows the consequences of choosing too small or too large threshold values. The clustering method is less sensitive to the parameters defining the minimum cluster size (T) and the maximum number of merges (M_{max}). In this study, they were kept in constant values.

TABLE I
EFFECT OF THRESHOLD PARAMETERS

parameter		number of actions		consequence
S_s	D	split	merge	
small	small	high	low	large number of clusters
small	large	low	low	bad classification accuracy
large	small	high	high	long computation time
large	large	low	high	small number of clusters

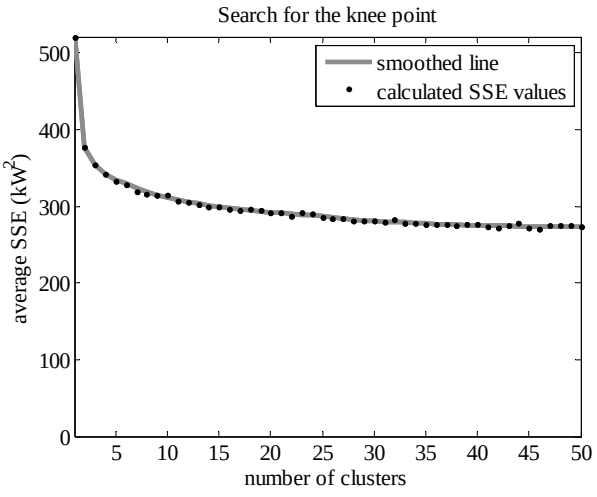


Fig. 7. Public administration load profile SSE as a function of number of clusters.

C. Accuracy Comparison

To verify the accuracy of the ISODATA clustering method, comparisons were made to alternative classification methods. Classification according to CIS customer class information and

allocation to the nearest existing customer class profile were selected as alternative classification methods. The accuracy of the individual load profiles was also verified. The forecasting capability of the load profiles was tested by comparing them with the actual measurements from the time period August 18, 2009 to December 31, 2009. Both expectation and standard deviation values were calculated for the load profiles, but only the load expectation values are studied in these comparisons.

1) Classification Method Comparison: The classification method comparison was made between five different methods: Previously presented ISODATA clustering, allocation to the nearest existing customer class profile and classification in three different accuracy levels according to the CIS customer class information. In this case, the customer class information in CIS is given with a three-digit number. The first number defines the customer's main customer class (e.g. industry), the second specifies classification further (e.g. metal industry), and the third gives the final customer class (e.g. manufacture of metal products). In the level 1 classification, only the first number was used and in levels 2 and 3 also, the second and third numbers were taken into account, respectively. After the classification, CIS-based customer class load profiles were calculated in the same way as the ISODATA-based customer class load profiles. The existing customer class profiles were provided by the Finnish Electricity Association (Sener) [3].

Fig. 8 presents the results for the accuracy comparison. It can be seen that the ISODATA clusters clearly have a smaller square sum of errors than the alternative classification methods, even though some of them had a larger number of customer classes (*c*). In Figs. 8 and 9, the average SSE is given to measurements normalized to 10 MWh/year energy consumption level.

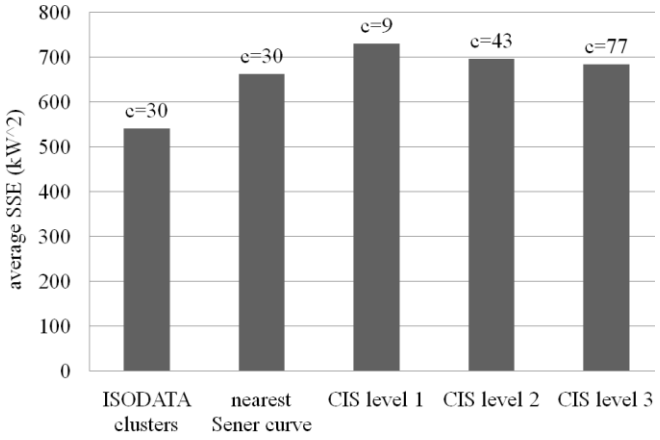


Fig. 8. Comparison of the classification methods.

2) Individual Load Profile Comparison: Here, the individual load profiles were formed based on the pattern vectors. In addition to previous calculations, standard deviations were also calculated for the pattern vector. The original temperature normalized measurements data was used in standard deviation calculation. Finally, the load profiles were formed by expanding each section in the pattern vector describing one week's consumption to cover the whole month.

The accuracy of the pattern vector based individual load profiles was compared to the accuracy of measurement based individual load profiles. The measurement based individual load profiles were formed directly from the previous year's measurements corresponding to the studied time period. In individual load profiling, the current practice in distribution companies is to use the previous year's measurements to model the electricity consumption in the current year.

Fig. 9 shows that pattern vector based load profiles produce better load forecasts than the load profiles formed directly from measurements. Holidays and the temperature dependency were taken into account in both studied load profiling methods. Fig. 9 also shows that load forecasts for the 92 outliers detected in Section IV-A are less accurate than load forecasts for the non-outliers.

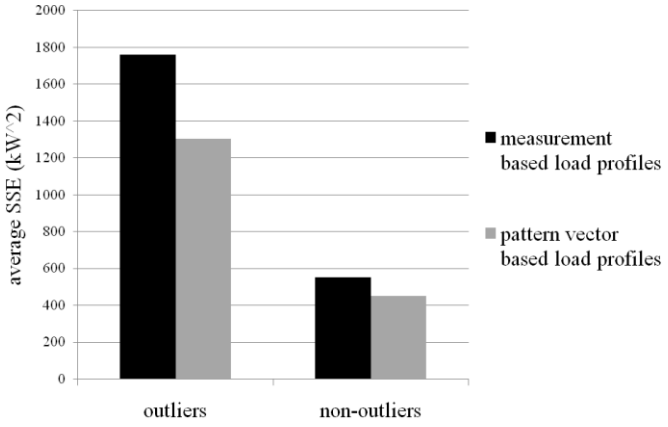


Fig. 9. Individual load profile accuracy comparison.

V. Discussion

The temperature dependency calculation, outlier filtering, clustering and load profile formation for all the 660 customers required approximately 60 seconds of CPU time (with 2.8 GHz Pentium 4 processor), not including the time used for the knee point search. In this paper, all the customers that passed the outlier filtering were subjected to clustering. In practice, the measurements can be compared with the existing customer class load profiles and only those customers that do not fit the existing load profiles can be subjected to clustering. This can reduce the computation time significantly.

It should be noted, that not all the customers should be clustered at the same time. For example, small residential customers should not be clustered simultaneously with large industrial customers. The clustering procedure is based only on expected load values and different sized customers have different standard deviations. Also, the load model accuracy requirements can be different. Large customers usually have lower stochasticity and thus better accuracy can be expected from their load models. Here this problem was solved by dividing the customers into six main customer classes. However, using the CIS information to divide the customers

into the main customer classes can cause new problems. Although rare, it is possible that some customers do not belong to the main customer class specified in CIS. Eliminating this problem would require an additional classification round where the classification of each customer is re-evaluated.

The final number of customer classes depends on how many sub-tasks the clustering is divided into and what is the desired number of clusters in each sub-task. Ultimately, the operator decides whether he wants to emphasize classification accuracy or to keep the number of customer classes easily manageable.

In the study above, only active power measurements were used. If reactive power measurements are available, the power factors can be taken into account in customer classification and load profiling.

Only customers with a limited amount of missing measurements were used in the clustering. The original measurement set also included measurements with long or frequent periods of missing data. Although the outlier filtering can be used to exclude these failed measurements from clustering, the missing data must be taken into account when forming individual load profiles for outliers. Handling these imperfect measurement series is a challenging task and should be a subject of further research. Also, possibilities to decrease the operator's role in customer classification should be studied.

VI. Conclusions

This paper presents an efficient method for the classification and load profiling of distribution network customers. The classification method utilizes AMR data, is based on ISODATA algorithm and involves temperature dependency correction and outlier filtering. The proposed method was implemented as a MATLAB program and tested with real measurement data. The results showed that the ISODATA algorithm can classify customers into well-separated clusters according to their electricity consumption data. It was also proven that the resulting customer classification is more accurate than the alternative classification methods: classification according to customer class information found in CIS and allocation to the nearest existing customer class profile.

Acknowledgment

The authors would like to thank the Satapirkan Sähkö Oy and its customers for providing the measurement data.

References

[1] G. Chicco, R. Napoli, and F. Piglione, "Comparison among clustering techniques for electricity customer classification," *IEEE Trans. Power Systems*, vol. 21, pp. 933–940, May 2006.

[2] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organising maps to classify electricity customers," presented at the *IEEE PowerTech Conf.*, Bologna, Italy, Jun. 2003.

[3] A. Seppälä, "Load research and load estimation in electricity distribution," Ph.D. dissertation, Helsinki Univ. of Tech., Espoo, Finland, 1996.

[4] G.J. Tsekouras, N.D. Hatziargyriou, and E.N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Systems*, vol. 22, no.3, pp. 1120–1128, Aug. 2007.

[5] Z. Zakaria, M.N. Othman, and M.H. Sohod, "Consumer load profiling using fuzzy clustering and statistical approach," in *Proc. 4th Student Conf. Research and Development*, Selangor, Malaysia, 2006, pp. 270–274.

[6] I.H. Yu, J.K. Lee, J.M. Ko, and S.I. Kim, "A method for classification of electricity demands using load profile data," in *Proc. 4th Annu. ACIS Int. Conf. Computer and Information Science*, Jeju Island, South Korea, 2005, pp. 164–168.

[7] B.D. Pitt and D.S. Kirschen, "Application of data mining techniques to load profiling," in *Proc, 21st IEEE Int. Conf. Power Industry Computer Applications*, Santa Clara, CA, 1999, pp. 131-136.

[8] S. Ramos and Z. Vale, "Data mining techniques application in power distribution utilities," in *Proc. IEEE/PES Transmission and Distribution Conf. and Expo.*, Chicago, IL, 2008.

[9] S.V. Verdú, M.O. García, F.J.G. Franco, N. Encinas, A.G. Marín, A. Molina, and E.G. Lázaro, "Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters," in *Proc. IEEE PES Power System Conf. and Expo.*, Atlanta, GA, May 2004, pp. 899–906.

[10] K.L. Lo and Z. Zakaria, "Electricity consumer classification using artificial intelligence," in *Proc. 39th Int. Universities Power Engineering Conf.*, Bristol, UK, 2004, pp. 443–447.

[11] D. Gerbec, S. Gašperič, I. Šmon and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," in *IEE Proc. Generation, Transmission and Distribution*, vol. 151, May 2004, pp. 395–400.

[12] N. Mahmoudi-Kohan, M.P. Moghaddam, and S.M. Bidaki, "Evaluating performance of WFA K-means and Modified Follow the leader methods for clustering load curves," in *Proc. IEEE/PES Power System Conf. and Expo.*, Seatle, WA, 2009.

[13] A. Mutanen, S. Repo, and P. Järventausta, "AMR in distribution network state estimation," presented at the *8th Nordic Electricity Distribution and Asset Management Conf.*, Bergen, Norway, Sept. 8–9, 2008.

[14] E. Lakervi and E.J. Holmes, *Electricity distribution network design*. 2nd ed. London, U.K.: Peter Peregrinus Ltd., 1995, ch. 11.3.

[15] M. Meldorf, *Electrical Network Load Monitoring*. Tallinn, Estonia: TUT Press, 2008.

[16] G.H. Ball and D.J. Hall, "ISODATA: a novel method of data analysis and pattern classification," Stanford Res. Inst., Menlo Park, CA, Tech. Rep. NTIS-AD-699616, Apr. 1965.

[17] C.W. Therrien, *Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics*. New York: Wiley, 1989.

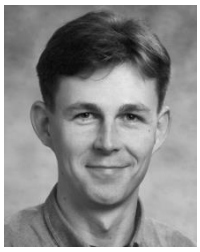
[18] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Symp. Mathematical Statistics and Probability*, Berkeley, CA, 1967, vol. 1, pp. 281–297.

Antti Mutanen was born in Tampere, Finland, on June 10, 1982. He received the M.Sc. degree in electrical engineering from Tampere University of Technology in 2008.

Currently, he is a Researcher and a Postgraduate student with the Department of Electrical Energy Engineering, Tampere University of Technology. His main research interests are load research and distribution network state estimation.

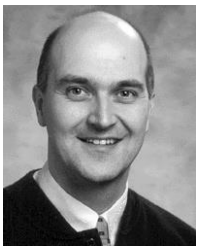
Maija Ruska received her M.Sc. degree in electrical engineering from Helsinki University of Technology in 2000.

Currently, she is a Research Scientist at the VTT Technical Research Centre of Finland. Her main interests are electricity and fossil fuel markets.



Sami Repo received his M.Sc. and Dr.Tech. degrees in electrical engineering from Tampere University of Technology in 1996 and 2001, respectively.

Currently, he is a University Lecturer with the Department of Electrical Energy Engineering, Tampere University of Technology. His main interest is the management of active distribution network including distributed energy resources.



Pertti Järventausta received the M.Sc. and Licentiate of Technology degrees in electrical engineering from Tampere University of Technology in 1990 and 1992, respectively. He received the Dr.Tech. degree in electrical engineering from Lappeenranta University of Technology in 1995.

Currently, he is a Professor in the Department of Electrical Energy Engineering, Tampere University of Technology. His main interests focus on electricity distribution and the electricity market.

Publication 6

A. Mutanen, P. Järventausta, M. Kärenlampi, and P. Juuti, "Improving distribution network analysis with new AMR-based load profiles," presented at the 22nd International Conference and Exhibition on Electricity Distribution (CIRED), Stockholm, Sweden, June 10–13, 2013.

Available at: https://doi.org/10.1049/cp.2013.0928

IMPROVING DISTRIBUTION NETWORK ANALYSIS WITH NEW AMR-BASED LOAD PROFILES

Antti MUTANEN
Tampere Uni. of Tech.
Finland
antti.mutanen@tut.fi

Pertti JÄRVENTAUSTA
Tampere Uni. of Tech.
Finland
pertti.jarventausta@tut.fi

Matti KÄRENLAMPI
ABB
Finland
matti.karenlampi@fi.abb.com

Pentti JUUTI
ABB
Finland
pentti.juuti@fi.abb.com

ABSTRACT

Automatic meter reading (AMR) is becoming common in many European countries. This paper shows how AMR measurements can be used to create new load profiles and how these new load profiles can be applied to improve distribution network analysis accuracy. In this paper, hourly electricity consumption data is used to update existing load profiles, cluster customers and create new cluster profiles, and specify individual profiles for selected customers, all of which are then used in distribution network analysis. The results between existing and new load profiling methods are compared. Comparisons are also made between different methods of AMR-based load profiling.

INTRODUCTION

With the advent of smart grids, the ways of operating distribution networks are changing. The amount of distributed generation (DG) is increasing and in order to accommodate the intermittent DG with reasonable network investments, automatic control of networks is increased. For example, demand response and coordinated voltage control are developed to keep the line flows and voltages within acceptable limits. All this tightens the requirements set for distribution network analysis. In smart grids, network planning and operation must be made more carefully in order to keep distribution networks within reduced operating margins. This applies not only to medium voltage (MV) but also to low voltage (LV) networks. Distributed generation and active network control are spreading also to LV side [1].

The timely and spatially correct commitment of the demand response and coordinated voltage control require accurate information about the state of the network [2]-[3]. It has been shown that load profiles have a big effect on the accuracy of distribution network state estimation [3], [4]. When forecasting the future states of the network, the load profiles have an even bigger role. State estimates and forecasts have a crucial role in network operation, especially in smart grids, and more accurate load models are needed to improve them.

Making customer level load models used to be expensive and time consuming, but now that automatic meter reading is quickly becoming common in many European countries, the effort required for load research has decreased considerably. Modern AMR systems provide abundant amounts of information on customer level electricity usage. This, along with the defects in existing load profiles [5], has motivated us to improve load profiling accuracy with AMR-based load profiles.

In Finland, distribution network customers are commonly classified to predefined customer classes, and the load of each customer is then estimated with customer class specific hourly load profiles. In an earlier publication [5] it was proven that in this environment a simple yet efficient method for improving load profiling accuracy is to update the existing load profiles with the help of AMR measurements. Even better results can be achieved if the load profile updating and customer reclassification are combined with the help of clustering methods. Also, creating individual load profiles can be beneficial, especially for the largest customers.

In this paper, we will present a revised version of the AMR-based load profiling method introduced in [5]. The load profiles calculated with this method will be compared with existing load profiles and measurements.

MATERIAL AND METHODS

In this study, we used hourly AMR measurements from two Finnish distribution companies; Koillis-Satakunnan Sähkö (Case 1) and Elenia Networks (Case 2). The measurements from Koillis-Satakunnan Sähkö were made between the 4th of December 2007 and the 3rd of March 2011. The starting time of each measurement varied and only those customers who had been measured for at least 13 months were selected for further analysis. 5343 such customers were found from the measurement database. The developed load profiling method requires measurement data from at least one year. The last month from the measurement data was reserved for the verification of results. From Elenia Networks, we had 7558 measurements done between the 10th of June 2010 and the 31st of October 2012. The last year from the measurement data was reserved for the verification of results.

Both measurement sets came from small towns and rural areas surrounding the towns. These measurements covered a wide variety of customer types ranging from small summer cabins to large industrial customers. In Case 1, the measurements were scattered across the

network operator's supply area in several municipalities. In Case 2, the measurements covered all the customers supplied by a substation feeding the town of Orivesi. For both cases, we had hourly temperature measurements and basic customer information. The original customer classification was known and network information enabled load flow calculations with original and new load profiles.

Figure 1 presents flow charts for the load profile updating and clustering methods used in this paper. After the measurements had been read and pre-processed, seasonal temperature dependency parameters were calculated for each customer using the method presented in [6]. The temperature dependency parameters were then used to normalize the measurements in to the long time average monthly temperatures. The temperature normalization was made so that measurements from several different years could be treated equally. Also, the normalized measurements were needed when the next year energy forecasts were made. If measurement data was available from several years, simple linear regression was used to forecast the next year's energy consumption.

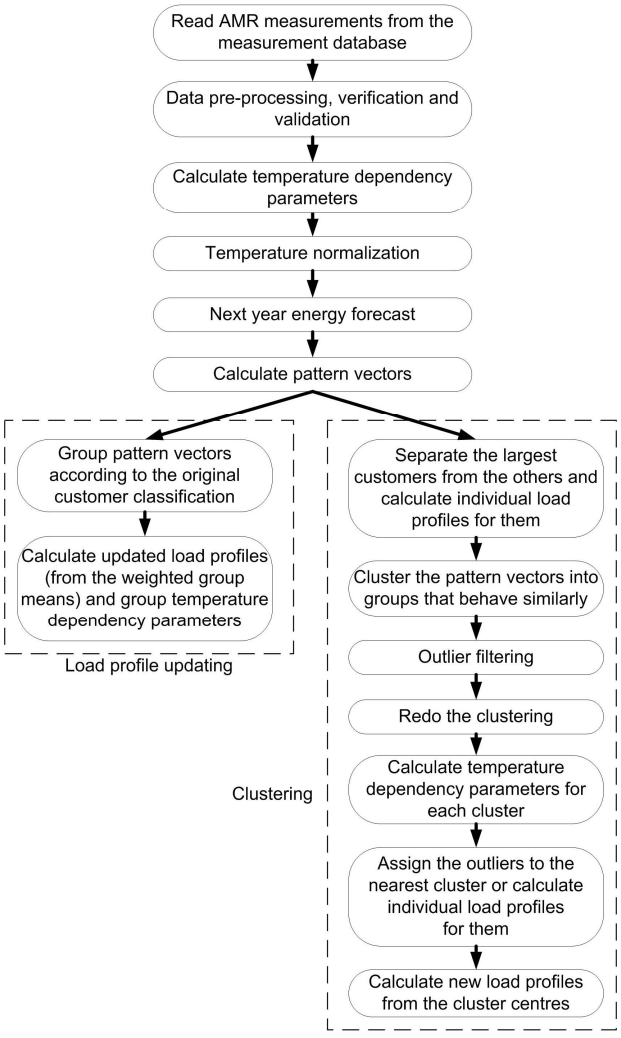


Figure 1. *Clustering and load profile updating methods.*

Pattern vectors describing the consumption of each customer were calculated from the normalized measurements. The pattern vectors consisted of 2016 values (12 months × 7 days × 24 hours = 2016) describing the average hourly consumption. Analysis of variance (ANOVA) was applied to determine if intraday behaviour on different weekdays was significantly different. If it was, then each weekday was modelled separately. If it was not, then all weekdays were modelled with a common weekday model.

At the beginning of the clustering procedure, the largest customers were separated from the others and individual load profiles were calculated for them. Then the pattern vectors were grouped into groups that behave similarly with the help of k-means clustering method. The original customer classification was used as a starting point for the clustering and pattern vectors were weighted according to the corresponding customer size (yearly energy). After this initial clustering, outliers were removed from the data. The customers with largest weighted distance from the cluster centres were selected for individual profiling and the customers with largest un-weighted distance were labelled as outliers and set aside (5 % of the total population). The clustering was redone and temperature dependency parameters for each cluster were calculated. Then the previously removed outliers were assigned to the nearest cluster and load profiles were formed from the cluster centres. Both the updated load profiles and cluster profiles were made compatible with the existing load profile format where each hour of the year has an expectation value and a standard deviation.

RESULTS

Case 1: Koillis-Satakunnan Sähkö

With the available AMR measurements, we were able to update 23 out of 38 customer class load profiles currently used in Koillis-Satakunnan Sähkö. Clear changes were observed in all the updated load profiles. Figures 2 and 3 show how the load profile for customer class 1 (housing) changed. From Figure 3, we can see that when the outdoor temperature is close to the average monthly temperature, the customer class sum load forecasted with the updated load profile matches to the measured sum load but when the temperature drops, the measured load exceeds the forecasted load. This is why we calculated temperature dependency parameters for each updated customer class. Temperature dependency information is especially useful when one is making short term load forecasts and temperature forecasts are available.

In distribution network analysis, one of the most important tasks is the forecasting of next year's peak loads. Temperature dependency information can help in this task; even it is not possible to make temperature forecasts so far ahead. Based on historical weather information, it is possible to determine a probable

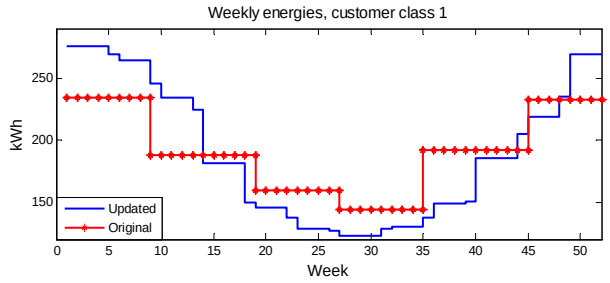


Figure 2. *Comparison of weekly energies in original and updated load profile.*

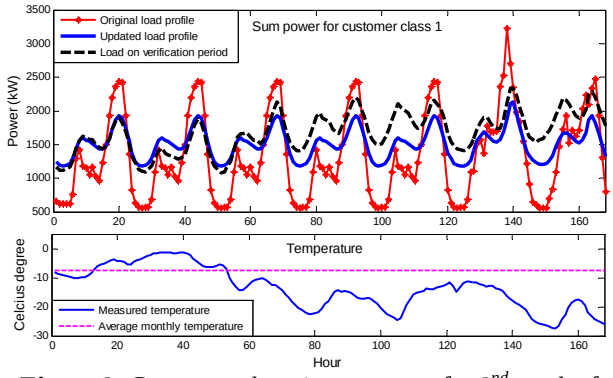


Figure 3. *Customer class 1 sum power for 2nd week of February.*

minimum temperature for a certain area to make "worst case" simulations. For areas studied in this paper, −25 °C was a good estimate for minimum daily temperature.

During the clustering phase, the customers were clustered in to 27 clusters and 100 individual load profiles were formed for large and abnormally behaving customers. The original customer classification was used as a starting point of the clustering but the final customer classification had little to do with the original customer classification. Only 15 % of the customers stayed in their original customer classes.

Since all customers are not (yet) measured with AMR and optimal clusters can be determined only for measured customers, the old and updated load profiles have to be used side by side with the cluster and individual profiles in network calculation. During this study, a modified prototype version of ABB MicroSCADA Pro DMS 600 -software was made to test this concept. The prototype software used all the aforementioned load profile types together. Old and updated load profiles were used for the unmeasured customers and cluster and individual profiles were used for the measured customers. Also, the operator could choose which load profiles to use. The prototype software was used first for LV network minimum voltage analysis but no clear differences between the load profiling methods were detected due to the stochastic nature of LV loads. The differences can be seen only when studying aggregated loads or when the sample size is large enough.

Table I shows average peak loads for all 5343 studied

Table I. *Comparison of peak load estimates on a customer level.*

Method	Average peak load (kW)		
	confidence level		
	50 %	90 %	95 %
Original load profiles	4.2	7.0	7.8
Updated load profiles	3.5	5.9	6.6
Updated load profiles -25 °C	4.1	6.4	7.1
Cluster profiles	3.8	5.8	6.4
Cluster profiles -25 °C	4.4	6.4	7.0
Peak load on a previous year	7.0		
Measured peak load on the verification period	7.17		

customers. When using 95 % confidence level, which is a typical confidence level when calculating peak loads, the original load profiles give too high peak load estimates but the updated load profiles and cluster profiles give good results when −25 °C minimum temperature is assumed (minimum temperature during the verification period was −26 °C).

Case 2: Elenia Networks

In Case 2, updated load profiles were calculated for 30 customer classes. As in Case 1, the updated load profiles gave significantly lower peak load forecasts than the original load profiles but when scaled to estimated yearly minimum temperature of −25 °C, the peak load forecasting accuracy improved.

In the clustering phase, the customers were clustered in to 30 clusters and 200 individual load profiles were formed for large and abnormally behaving customers. With the updated load profiles, the verification period square sum of forecasting errors decreased 38 % when compared with the original load profiles. With the cluster profiles this value was 57 %.

Tables II and III show verification period peak load forecasts calculated on a distribution transformer level (i.e. sum of all the customers supplied by the specific transformer) and on a substation level. On average, the best distribution transformer level peak load forecasts

Table II. *Comparison of peak load estimates on a distribution transformer level.*

Method	Average peak load (kW)		
	confidence level		
	50 %	90 %	95 %
Original load profiles	44.7	57.9	62.0
Updated load profiles	36.6	44.9	47.5
Updated load profiles -25 °C	47.8	55.9	58.4
Cluster profiles	39.1	46.2	48.6
Cluster profiles -25 °C	50.5	57.4	59.7
Peak load on a previous year	56.8		
Measured peak load on the verification period	53.7		

Table IV. *Comparison of peak load estimates on a substation level.*

Method	Peak load (MW)		
	confidence level		
	50 %	90 %	95 %
Original load profiles	17.3	17.8	17.9
Updated load profiles	15.1	15.3	15.4
Updated load profiles -25 °C	19.8	20.0	20.1
Cluster profiles	15.0	15.2	15.2
Cluster profiles -25 °C	19.8	19.9	20.0
Peak load on a previous year	19.9		
Measured peak load on the verification period	19.3		

were achieved using updated load profiles and 90 % confidence level. Also, the original and cluster profiles provided good results with 90 % confidence level. The selection of the best confidence level proved to be difficult since for small distribution transformers with few customers the 95 % confidence level provided the best results but for large distribution transformers with many customers the 50 % confidence level was the best. On the substation level peak load forecasts the effect of used confidence level was small and the selected minimum temperature dictated the peak load forecast magnitudes. In Case 2, the forecasted peak loads were systematically higher than the actual measured peak loads since there was a 6.8 % drop in the electricity consumption between the load profile identification and verification years. This drop could not be explained entirely with load temperature dependency and was probably caused by economic factors which were not taken into account in this study.

CONCLUSIONS

This paper presented two alternative methods for calculating AMR based load profiles. The first method used AMR measurements to update the existing customer class load profiles but kept the customer classification unchanged, while the second method used k-means clustering to update both the load profiles and customer classification. Also, individual load profiles were formed for large and abnormally behaving customers. Both the presented load profiling methods modelled the load temperature dependency and random variation separately.

Load temperature dependency information is especially useful when one is making short term load forecasts but it can be used to improve next year peak load forecasts as well. In cold countries, the peak loads occur during the coldest days of the year and it is quite easy to determine a suitable peak load calculation temperature from the historical temperature information.

The new AMR based load profiles were clearly better than the original load profiles. When forecasting future loads, the cluster profiles had the best average fit but no significant improvement in peak load forecasting capability was detected when compared with the updated load profiles.

Although the results were better than with the original load profiles, the customer and distribution transformer level peak load forecasting proved to be a challenging task even for the new AMR based load profiles. Since the previous year's peak load seems to give a good indication for future peak loads, the direct usage of AMR measurements in distribution network peak load calculation should be studied. Also, the possibility of using distribution transformer level load models, instead of aggregated customer level load models, in MV network calculation could be studied.

REFERENCES

[1] S. Repo, D. Della Giustina, G. Ravera, L. Cremaschini, S. Zanini, J. M. Selga and P. Järventausta, 2011, "Use Case Analysis of Real-Time Low Voltage Network Management," presented at the *2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe)*, Manchester, UK.

[2] A. Kulmala, S. Repo and P. Järventausta, 2009, "Increasing penetration of distributed generation in existing distribution networks using coordinated voltage control," *Int. Journal of Distributed Energy Resources*, vol. 5, 227-255.

[3] M. Biserica, Y. Besanger, R. Caire, O. Chilard and P. Deschamps, 2012, "Neural Networks to Improve Distribution State Estimation – Volt Var Control Performances," *IEEE Transactions on Smart Grid*, Vol. 3, No. 3.

[4] A. Mutanen, S. Repo and P. Järventausta, 2008, "AMR in Distribution Network State Estimation", presented at the *8th Nordic Electricity Distribution and Asset Management Conf.*, Bergen, Norway.

[5] A. Mutanen, S. Repo and P. Järventausta, 2011, "Customer Classification and Load Profiling Based on AMR measurements," presented at the *21st International Conference and Exhibition on Electricity Distribution*, Frankfurt, Germany.

[6] A. Mutanen, M. Ruska, S. Repo and P. Järventausta, 2011, "Customer Classification and Load Profiling Method for Distribution Systems," *IEEE Transactions on Power Delivery*, Vol. 26, No. 3.

Publication 7

B. Stephen, A. Mutanen, S. Galloway, G. Burt, and P. Järventausta, “Enhanced load profiling for residential network customers,” IEEE Transactions on Power Delivery, vol. 29, no. 1, Feb. 2014.

Available at: https://doi.org/10.1109/TPWRD.2013.2287032

Enhanced Load Profiling for Residential Network Customers

Bruce Stephen, *Member, IEEE,* Antti J. Mutanen, Stuart Galloway, Graeme Burt, *Member, IEEE* and Pertti Järventausta, *Member, IEEE*

***Abstract*— Anticipating load characteristics on low voltage circuits is an area of increased concern for Distribution Network Operators with uncertainty stemming primarily from the validity of domestic load profiles. Identifying customer behavior makeup on a LV feeder ascertains the thermal and voltage constraints imposed on the network infrastructure; modeling this highly dynamic behavior requires a means of accommodating noise incurred through variations in lifestyle and meteorological conditions. Increased penetration of distributed generation may further worsen this situation with the risk of reversed power flows on a network with no transformer automation. Smart Meter roll-out is opening up the previously obscured view of domestic electricity use by providing high resolution advance data; while in most cases this is provided historically, rather than real-time, it permits a level of detail that could not have previously been achieved. Generating a data driven profile of domestic energy use would add to the accuracy of the monitoring and configuration activities undertaken by DNOs at LV level and higher which would afford greater realism than static load profiles that are in existing use. In this paper, a linear Gaussian load profile is developed that allows stratification to a finer level of detail while preserving a deterministic representation.**

***Index Terms*— Automatic meter reading (AMR), domestic load profiling, energy demand, low-voltage (LV) networks.**

I. Introduction

THE low-voltage (LV) network and the consumers on it has been a relative unknown quantity in power system design and operation with highly generalized profiles of domestic households being used to make decisions in all but a few exceptional cases [1]. The advent of smart metering has the potential to change much of that but with the increased volumes of household energy use data comes questions on how best to employ it and prior to that how to understand it in the first place. It has been postulated in smaller scale studies that domestic customers can be profiled according to energy usage time and magnitude. How these profiles aggregate together on a low voltage feeder is of interest to distribution network operators (DNOs) who traditionally would assume load was merely a multiple of a single homogenous domestic profile – Fig. 1 shows how this is not necessarily the case. Even on similar dwellings the customer behavior can be very diverse.

As some of the key technologies of smart grids are realized, the concerns regarding legacy infrastructure become more apparent. Increasing penetrations of micro-generation are challenging the usefulness of this assumption as excess domestic generation tips residential feeders into reverse power flows. While generation such as photovoltaic can be predicted to some degree of accuracy, there needs to be further work on modeling the loads that absorb them. Behavioral factors are identified in [2] that influence the load profile breaking energy demand into 2 root causes: behavioral determinants – habit driven, relatively flexible; and physical determinants – driven by environmental factors and building design. Behavioral drivers are the one which invoke most variability, [3] noted in an overview of advanced tariffs (e.g. real time pricing) that not all customers could be suited to these; demographics such as young families – no flexibility, constant temperature and the elderly who also require constant temperature. Then there are those who maintain a constant load already with the only losses stemming from dwelling disrepair/insulation shortcomings (cf. the "physical determinants" of [2]).

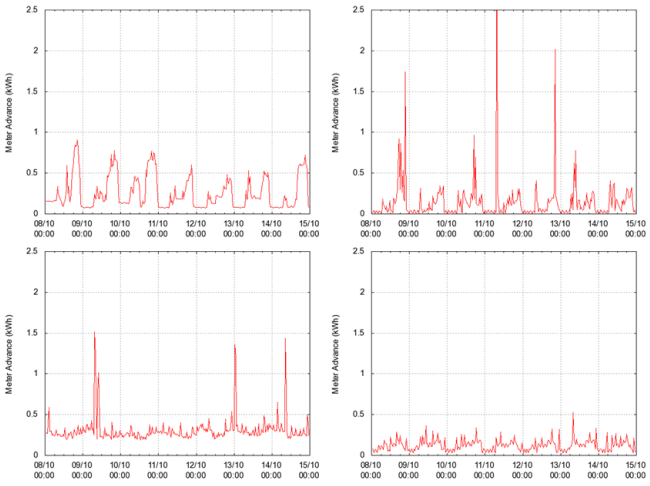


Figure 1. The 30-min resolution residential loads over a single week from similar dwellings.

B. Stephen is with the Advanced Electrical Systems Research Group, Institute of Energy and Environment, University of Strathclyde, Glasgow, G1 1XW, U.K. (e-mail: bruce.stephen@strath.ac.uk).

A. J. Mutanen is with the Electrical Energy Engineering Department, Tampere University of Technology, Tampere FI-33101, Finland (e-mail: antti.mutanen@tut.fi).

S. Galloway is with the Electronic and Electrical Engineering Department, University of Strathclyde, Glasgow G1 1XW, U.K. (e-mail: s.galloway@eee.strath.ac.uk).

G. M. Burt is with the Institute for Energy and Environment, University of Strathclyde, Glasgow G1 1XW, U.K. (e-mail: g.burt@eee.strath.ac.uk).

P. Järventausta is with the Electrical Energy Engineering Department, Tampere University of Technology, Tampere 33101, Finland (e-mail: pertti.jarventausta@tut.fi).

With consumer technology acquisition at its highest ever level, and expected to continue to grow, such profiles can only become invalid quicker thus reinforcing the case for data driven methodologies to be used. In this paper, an alternative representation of domestic load is considered, that of a composition of usage levels strata generated dynamically from Smart Meter data. Embedding this representation in a probabilistic model allows a quantifiable comparison to be made between profiles generated by different dwellings and how these can change. This paper will present a framework for analyzing the consumption habits of domestic energy customers which will be illustrated through the application to actual half hourly metered properties.

II. Residential Loads

The absence of low voltage metering means that until recently very little knowledge exists on the low voltage customer's true load profile. This section reviews some of the current practices and looks at how larger loads are dealt with on the medium voltage (MV) network.

A. Current Profiling Practice

The current practices tend to involve metering relatively small samples of households and then averaging over these. The following outlines examples from the U.K. and Finland.

1) United Kingdom: For the U.K., it was decided in the mid-1990's that to facilitate market operation, 8 load profiles would be used to represent the types of customers on the network. Of these profiles, Profile Class 1 [4] is the only one that represents the residential customer unconstrained by usage times. The form of the profile is 48 half-hourly usage levels that correspond to the market settlement periods for every settlement day in a year. These are developed from recruited sample households with hi-resolution meters; homes in the samples for the 14 U.K. grid supply points are selected from rule-based stratifications (high medium low) of annual consumption obtained from retail billing. Averages of the half-hourly data are weighted by the proportions of the population at a given grid supply point in a given strata, yielding a load profile that takes the form of a 48×365 matrix.

2) Finland: Finnish electric utilities started to co-operate in load research in the 1980's and in 1992 Finnish Electricity Association (FEA) published customer class load profiles for 46 different customer classes, 18 of which are for housing and the rest for agriculture, industry and services. The housing profiles are further divided by dwelling-type, heating solution and major appliances. Each load profile contains expectation and standard deviation values for every hour of the year [5]. Although old, the FEA load profiles are still the only publicly available load profiles. The most prominent shortcoming of these profiles is their age; during the past 20 years electricity consumption has experienced significant changes, the amount of heat pumps and air-conditioners has multiplied, the use of entertainment electronics has increased and electricity consumption in recreational dwellings has changed [6]. Furthermore, in the future, the changes will be even bigger if plug-in hybrids, customer-specific distributed generation and demand response activities become popular. The load profiles also suffer from small sample sizes, short measurement periods and errors caused by geographical generalization. The load profiles are created to model the average Finnish electricity consumption. They do not take into account the regional differences in electricity consumption, which originate from different climate conditions and socioeconomic factors. Consequently, the strategies used are error prone: the type of the customer is usually determined through a questionnaire when the electricity connection is contracted and then rarely updated. In reality, the customer type may change, for instance, because of a change in the heating solution, an addition of new devices, such as air conditioning or the change of customer activity (e.g. from agriculture to pure housing).

B. Related Load Profiling on MV Network

In [7], Probabilistic neural networks (PNNs) were used to assign consumers to load profiles – these are closely related to a Parzen Window and essentially smooth input data into a probability density function (PDF) of observations. 10 load profiles resulted but different cluster validity measures resulted in conflicting optimal number of clusters. An assortment of clustering techniques are used in [8] on 234 non-residential customers metered on the MV network at 15-min intervals with the objective of grouping them into a small number of classes for tariff formulation. Reference [8] noted that theoretically robust means of choosing the number of clusters would be required as conflicts between cluster validity criteria could arise [7]. Techniques used include hierarchical clustering (with Euclidean distance), self-organizing maps, K-Means and Fuzzy K-Means. Dimensionality reduction of the 96-dimensional space into a more manageable subspace was also performed allowing the 'informative' hours/periods to be identified. ISODATA (Iterative Self Organizing Data Analysis Technique) was used in [9] to cluster industrial customers into load profile classes; outliers in training data were defined as customers with high intra-day variation and customers with high monthly variation were discarded.

Although load profiling on the MV network has received attention, the criteria associated with it are not the same; it was noted in [9] that large customers tend to have a small standard deviation in their load and hence produce a more accurate load profile lessening the need to encode variability in the profile representation thus emphasizing the need to encode variability in the smaller residential customer profiles as outlined in [10].

III. AMI/AMR Status

A number of countries are committed to upgrading their housing stock to AMR systems or smart meters. In the U.K. and Finland, large electricity customers are already metered on half hour or hourly basis but the state of domestic smart metering is different [11].

In Finland, full smart meter roll-out is currently underway and a significant number of meters have already been installed [11]. Legislation requires electricity distribution network operators to equip at least 80 % of their customers with hourly

metering by the end of the year 2013. Daily meter reading, support to demand response, and outage registration are also required [12]. One novel feature in Finnish AMR installations has been to integrate AMR system with control center applications of SCADA and distribution management system (DMS) in order to use AMR meters in real-time low-voltage network management and fault indication [13].

For the U.K., AMR will provide advance data at a 30-min resolution, most likely communicated at the end of a 24-h period. Full scale roll-out is scheduled to begin in 2014 and finish in 2019 although some crucial parts of the program, such as details concerning national data and telecommunication services, are yet to be decided [14].

IV. Residential Profiling Requirements

Reference [15] identifies that "individual consumer behavior and their everyday practices accounts for a substantial proportion of household energy consumption". In identical houses, it was noted that this can vary by up to 300–400% as a result. The drivers for variability are multi-factorial: [16] identifies that different socio-economic types will contribute different amounts to energy demand using the local area resource access model (LARA) – high levels of socioeconomic and geographical disaggregation were noted in the U.K. Although the credit rating agency groups were noted, [16] uses U.K. output area classification (OAC) to segment U.K. households into seven groups with different socio-demographic characteristics with largely self explanatory labels (e.g. "Blue Collar communities", "City Living", "Countryside", "Prospering suburbs". A "Culture based approach to behavior" is explored in [17] by identifying energy usage behaviors as a means of finding opportunities to invoke changes in behavior. In [17], the "Energy Cultures" framework was proposed to explain different causal facets of energy use which can be summarized as: Material Culture which is characterized by: insulation, heating devices and influenced by: Regulation, income, available technology; Cognitive norms which are characterized by: social aspiration, tradition, environmental concern and influenced by: Education, upbringing, demographics; Energy Practices which are characterized by: Number of rooms, Maintenance of technology and influenced by: Social Marketing, Energy Price Structure. As discussed, load profiles for the residential customer have been largely homogenous arrangements that were calendar based rather than behavior driven. With AMI/AMR/Smart Metering measurements providing extensive and detailed load and resulting variability, a representation is needed to capitalize on this and provide utility stakeholders with the information they require to increase reliability and efficiency. Regarding actual behavior, it is highly unlikely that all residential customers behave the same, so the representation must be able to accommodate a finite number of heterogeneous behaviors and do so in a compact manner thus enabling the representation to be utilized without unfeasibly large computing resources. For each heterogeneous behavior encountered, the traditional quantity of interest is the expected value of load; time of use is the other traditional concern so what is really required is a coupling of time of use with load magnitude. AMI in the U.K. and Finland provides data with half hour or 1-h resolution allowing this quantity to be represented as a discrete vector rather than a functional approximation. Where curve fitting or regressive approaches may not suffice is in the provision for capturing load variability – the confidence with which a given load's expected value is expressed is also necessary. For forecasting purposes, which may arise in highly localized power systems, the relation between time of day loads can inform a short term forecast (weather related behavior change). Detection of anomalous behavior is another requirement that would provide indication of fault condition or, over longer terms, new classes of customer emerging (e.g. greatly reduced loads through adoption of storage or uptake of more efficient appliances). Additionally, the capture of changes in behavior should be allowed through the representation.

V. Load Model Design

A. *Load Probability Distributions*

In load research, electric loads are often assumed to have a Gaussian distribution even though this is not the case. Previous studies [18]–[20] have tried to find the best probability distribution to model electric load behavior. In these studies, beta, gamma, and log-normal distributions have been found to model electrical loads better than Gaussian distribution. Fig. 2 shows that, when scored with Bayesian Information Criterion (BIC) [21], the log-normal distribution best describes U.K. residential loads out of several candidate probability distributions and is significantly better than the normal distribution. Also, by log-normalizing the data, it can be transformed to behave like a Gaussian distribution, which, in turn, enables the use of algorithms designed for the more tractable Gaussian distribution.

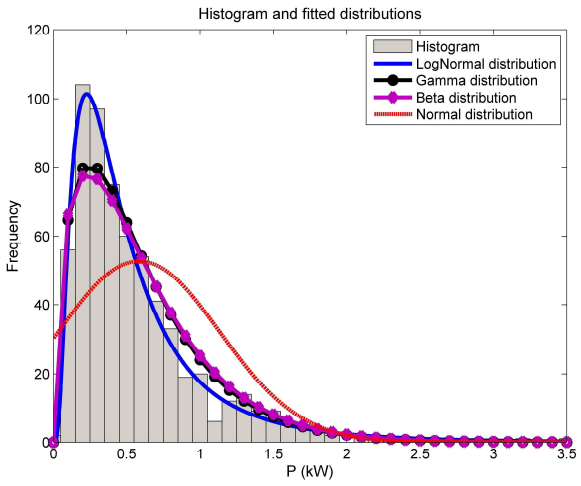


Figure 2: Histogram and fitted distributions for half-hour period 15:00-15:30 in January (weekdays only).

B. *Expressing Uncertainty Through Probabilistic Models*

The general form of models proposed in this paper is one of a non-stationary multivariate Gaussian distribution over 48 half-hourly advance periods. In [20], it was noted that variability of even a single customer is such that an individual load pattern

cannot be obtained – thus the importance of modeling the distribution rather than (just) the expected value. This section discusses several model families that may be used to express multimodality and dependence and in such a way that the representation maintains its compactness.

1) Mixture Models: A finite mixture model permits an arbitrary probability distribution to be approximated by a linear combination of weighted likelihoods drawn from a set of simple parametric distributions:

$$P(x) = \sum_{i=1}^{M} \pi_i P(x; \theta_i) \tag{1}$$

If this were a Gaussian mixture model, then the components would be Gaussian parameterized as follows:

$$P(x) = \sum_{i=1}^{M} \pi_i P(x; \mu_i, \sigma_i^2) \tag{2}$$

where x is the observation variable, θ_i is the parameter vector for the i^{th} distribution, π is the vector of mixing weights and M is the number of distributions used to approximate the implied observation distribution.

2) Factor Analysis: As daily meter advances are represented as a 48 dimensional vector here, it is difficult to assess which times of use influence each other and how. Multivariate data can sometimes contain correlation between variables that are so strong, these can be amalgamated allowing only the most informative or uncorrelated variables to be represented in a space of reduced dimensionality. Two examples of models which can reduce the dimension of an observation space and thus discard uninformative variables and reveal dependency structure are Principal Component Analysis (PCA) [22] and Factor Analysis [23]. PCA is based around the eigenvectors that correspond to the eigenvalues of the covariance matrix of a multivariate observation. Factor Analysis assumes a linear mapping between such an observation space x and its lower dimensional representation z:

$$x = \mu + \Lambda z + u \tag{3}$$

where Λ is the factor loading matrix that transforms observation x into a lower dimensional representation z. μ is the mean of the observation variable. Ψ is a diagonal covariance matrix attached to the zero mean distribution from which Gaussian noise u is drawn.

$$u \sim N(0, \Psi) \tag{4}$$

Factor Analysis does not impose the constraint of a common variance for all features and furthermore has a probabilistic model associated with it in the form of a multivariate Gaussian

$$P(z) = N(0, \Lambda\Lambda^T + \Psi) \tag{5}$$

Owing to the linear Gaussian semantics of the model, the observation space is also assumed to be Gaussian

$$P(x|z) = N(\mu + \Lambda z, \Psi) \tag{6}$$

where Λ is of particular use as interpretation of its rows/columns reveals the relations between variables in the observation space.

3) Mixtures of Factor Analyzers: For the situation where sub-populations exist in the observed data and multivariate dependency is non-homogeneous, the factor analysis model may be embedded in a mixture model [24].

$$P(x) = \sum_{i=1}^{M} \pi_i P(x; \mu_i + \Lambda_i z, \Psi_i) \tag{7}$$

Extending the mixture model to factor analysis, allows multiple sub-populations in a sub-space to be captured. The mixture of factor analyzers (MFA) model is particularly appealing to the load profiling application as it encodes not only the broad customer behaviors in the form of the model means but also expresses the variability over a day in a compact parameter set which also relates the advance times in terms of their variability.

C. Parameter Estimation and Model Order Selection

Beginning with a set of smart meter data there are two stages to go through before a model can be obtained: model selection and parameter estimation. Model selection decides on the cardinality of the model, the number of mixture components and the number of factors in the case of the Gaussian Mixture and MFA models previously discussed. Optimization techniques that estimate the parameters of statistical models from exemplar data are often based around maximum likelihood estimation (MLE). Model order selection techniques often require parameters for a set of models to be learned then the optimal one chosen using some likelihood-based measure such as BIC or Akaike information criterion (AIC):

$$AIC(X, \theta) = -2N \sum_{n=1}^{N} \log P(x_n|\theta) + 2M \tag{8}$$

These select the most likely number of parameters M while penalizing overly complex models of a data population of size N. Model complexity can harm the generalization capabilities of a model by encoding too many specific eventualities in it.

While more complex parameter estimation techniques exist such as Monte Carlo-based methods and variational inference, for illustrative purposes, the simpler maximum likelihood estimate-based formulation of the Expectation Maximization algorithm [25] can be used on both the mixture models and the factor analyzers.

VI. Learned Residential Load Profiles

To illustrate the models proposed in this paper, load models are learned for a group of 32 residential customers. Since load behavior is seasonal, separate load models are formed for each month. In the following examples, only January's load models are shown.

A. Gaussian Mixture Load Model

Using the January meter data for 32 residential properties, 50 Gaussian Mixture Models (GMM) were learned using maximum likelihood EM; from these 50 the optimal number of

mixtures was selected using BIC, the results of which are shown in fig. 3. Fig. 3 demonstrates a pronounced minimum at 16 components but also reveals some important features of the data; the asymptotic behavior of the left-most extreme indicates that a single Gaussian distribution provides the poorest fit to the data which reinforces the need to provide for multimodal behaviour. Furthermore, a large number of behaviours does not adequately represent the behaviour of residential customers either – domestic loads would appear to have, as far as a Gaussian representation is concerned, a relatively small number of plausible forms, although as stated in the outset, not a single one.

One advantage of the mixture model over say a neural network-based clustering approach such as a self organizing map is that an element of determinism can be obtained through inspection of the parameters. Fig. 4 shows the component means for the optimal parameterized GMM load model. This demonstrates the recurring load profile forms found in the 32 residential properties over the January period. One limitation of the Gaussian Mixture Model load profile is that owing to the high dimensionality of the data, it has difficulty expressing the dependence between advance times present in residential loads.

B. Mixture of Factor Analyzers Load Model

For an MFA mixture, an additional consideration is added to the model selection process in that one can trade off between mixtures (which accommodate various expected load profiles) and subspace dimensions (which capture the drivers of the correlation and variance structure).The MFA models offer even further insight into the nature of the load profiles discovered. Full covariance structure can be obtained for all mixture components regardless of the dimensionality of the data or the sparseness of the subpopulation that forms a mixture component. A covariance matrix can be reconstituted from the factor loading matrix as shown in (5), an example of such a covariance matrix is shown in Fig. 5 as a heatmap representation: this shows how meter advances across the 48 daily intervals influence each other for a given load profile. Dark red areas are strong positive correlations i.e., when a given (row) advance increases, the corresponding (column) advance increases. Blue areas show negative correlation – increases in (row) advance size result in decreases in corresponding (column) advance. The 48 dimensional representation can pose difficulties in articulating in the relationships between advances due to the high dimensionality of the data [26]. The additional advantage of the MFA model is that the factor loading matrix yields a representation of dependence between dimensions as a vector plot in the low dimension subspace. Fig. 6 shows one example of this from a single component. The vectors that correspond to each advance can be interpreted as follows [27]: The arrows are the eigenvectors of a covariance matrix with relative directions representing their implied linear dependence: alignment is high correlation while opposition is high negative correlation. Right angles imply linear independence. It should be noted here that correlation i.e., linear dependence is being modeled, this does not necessarily indicate the presence or absence of non-linear dependence – the MFA model approximates non-linear dependencies with piecewise linearity. In the example in Fig. 6, advances at time periods 45–47 (10 P.M. to 11:30 P.M.) show a strong correlation reflecting late evening habits with

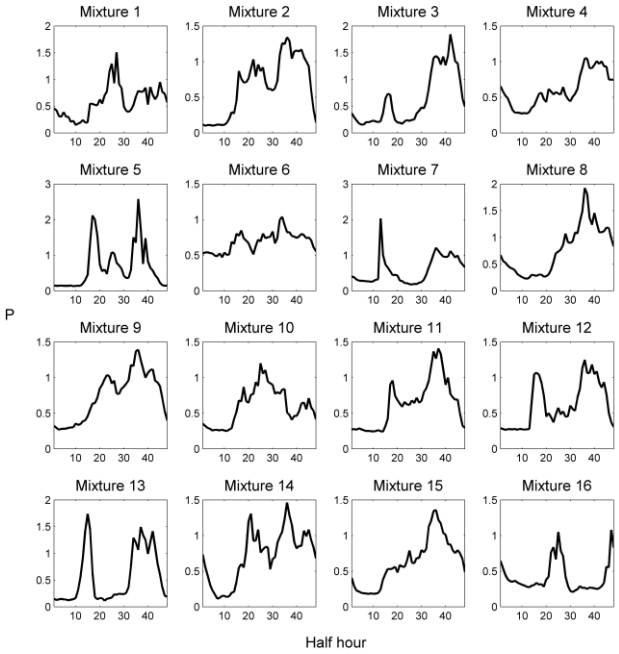


Figure 4: The 16 profile means found by the Gaussian Mixture.

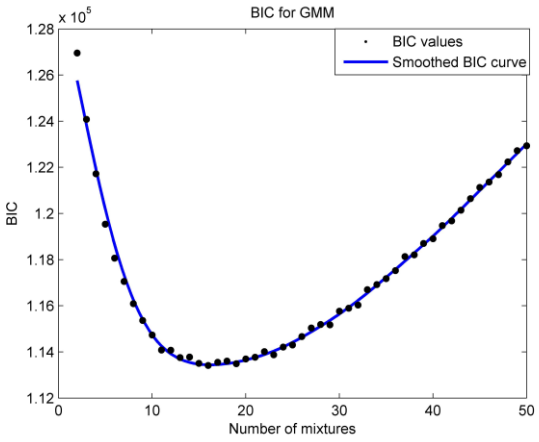


Figure 3: Selection of the optimal number of customer profiles a GMM load model should represent.

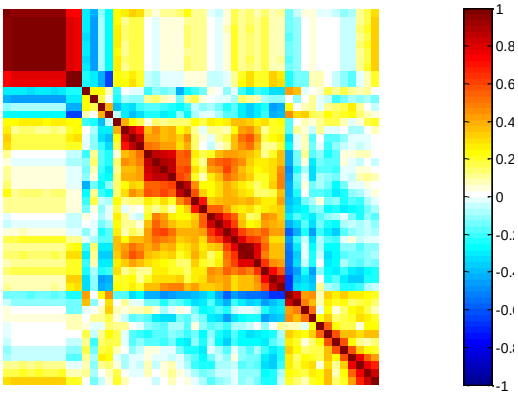


Figure 5: Example covariance matrix from one component of a GMM. Note the very strong correlations for the advances in the early hours of the morning.

little temporal variation and duration in the order of hours. Similar dependence structures are exhibited during the early hours of the morning as Fig. 5 demonstrates.

VII. Results and Practical Considerations

This chapter shows how the above presented load models could be used in practice and compares their performance to existing load models.

A. Load Model Allocation

Before the learned load models can be used, they must be compiled into customer specific monthly load profiles. January's load profile for all 32 customers can be compiled from the 16 previously learned day models, all we need to do is to find out which models best describe the customer's behavior on each day of the week. As an example, Fig. 6 shows how the Gaussian mixture load models are allocated for 4 different residential customers. Customer 17 shows remarkably consistent behavior, exhibiting the same profile for both weekday and weekend usage. Customer 29 switches between multiple profiles although does sometimes remain in the same one for more than one day. Customer 5 exhibits a near perfect separation in weekday/weekend electricity usage while Customer 31 switches between 3 profiles, always exhibiting the same energy usage characteristics on a Sunday. A single Gaussian distribution is not enough to describe a customer's behavior on each day of the week, so the final load model is constructed as a weighted average over all the mixtures in the model. This weighting is performed according to the occurrence counts of particular mixtures/profiles seen for a given customer during the period over which the training data was collected.

B. Comparison to Existing Load Models

In order to verify the accuracy of the proposed load modeling methodology, a comparison is made between the current British load modeling method (standard load profile), GMM and MFA. February's load forecasts are created using these methods and the forecasts are then compared to the real measured values. Since we have measurement data from only one year, the GMM and MFA model parameters are learned from January's data while February's measurements are reserved for verification. The selected Standard Load Profile (SLP) corresponds to the geographical location and type of the studied loads (domestic unrestricted customers). Both the GMM and MFA models are constructed using 16 mixtures. With 16 mixtures, the AIC for MFA model is lowest with ten subspace dimensions. For comparison, a MFA model with two dimensions is also built. The load forecasts were scaled to match the estimated energy consumption in February.

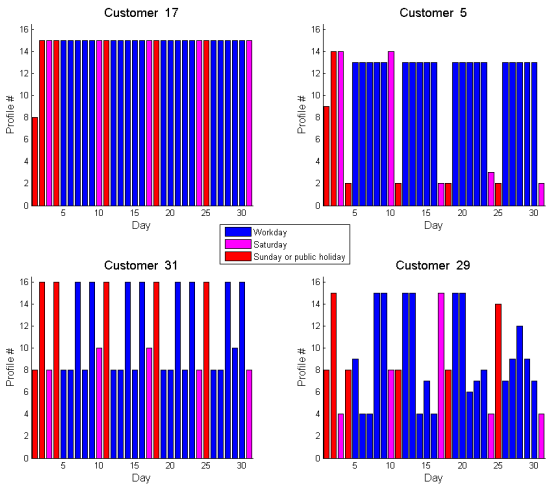


Figure 6: Demonstration of the daily variability of four residential customers with respect to day of the week.

C. Load-Flow Calculation

In practical applications, it is often important to estimate maximum (peak) or minimum (valley) loads. This is where the models of load variability are needed. When we know the load variability, we can calculate peak or valley loads with different confidence levels. In Finnish network calculation, 95% confidence is typically used when calculating maximum line flows [28].

*1) Simulation Network: T*he simulation network is based on a test network presented in [29]. Only the LV part of the test network is modeled in this study. The feeding MV network is modeled with a voltage source with 90 MVA short circuit power. The model incorporates a 500 kVA, 11 kV/433V ground mounted distribution transformer and four LV feeders each supplying 96 domestic customers. One LV feeder is modeled in detail and the other three are modeled as lumped loads, as shown in Fig. 7. The LV feeder is 300 meters long, it comprises two segments of cable, 150 m of 185 mm^2 and 150 m of 95 mm^2 cable. Single-phase customer connections are distributed evenly along the feeder and are connected to the main feeder with 30 m long 35 mm^2 service cables. Load points of phase L1 are populated with real metered data.

2) Simulation Results: Statistical load flow was performed on the simulation network. Since there is no explicit method for summing log-normally distributed variables, the following simplification was made when summing loads during the load flow calculation: Expectation values and variances were calculated for the log-normally distributed loads, expectation values and variances were then summed and log-normal distribution parameters were recalculated as in [30]. Load flow was calculated for every half hour of February using three different load profiles: SLP, GMM and MFA based load profiles. With GMM and MFA models, 95% confidence level was used. Maximum line currents and minimum node voltages were calculated and compared with the values calculated with real measured loads. Fig. 8 shows the estimated and "measured" maximum currents and minimum voltages on the phase L1 of the simulation network main feeder. The current and voltage values achieved with GMM and MFA models are very close to the real maximum and minimum values. Designing or operating the LV network based on Standard Load Profiles would be difficult since they do not take the peak or valley load situations into account correctly. GMM and MFA models were superior compared to SLP model even though January's load models were used to forecast February's load. More accurate models could have been created if measurements from the previous February had been available.

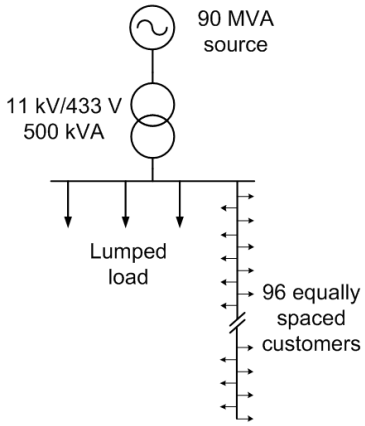


Figure 7: Single line diagram of the simulation network.

Euclidean distance, Peak & Valley estimates and Peak & Valley estimates with 95% confidence, were calculated for both aggregated load estimates and their corresponding actual values; this comparison is shown in Table I. With GMM and MFA (2D) models, the smaller Euclidean distance demonstrates they track aggregated load significantly better than the ones calculated with SLP. The MFA (10D) had a poor fit when evaluating performance with Euclidean distance, which may be down to overfitting of the covariance matrices in the higher dimensional space.

TABLE I
ACCURACY METRICS FOR DIFFERENT LOAD MODELS

Criteria	SLP	GMM	MFA (2D)	MFA (10D)
Euclidean distance	74.11	69.58	70.22	74.12
Peak estimate (real 23.1 kW)	19.32	18.62	18.69	18.26
Peak estimate with 95 % conf. interval	-	22.69	23.20	22.73
Valley estimate (real 2.93 kW)	3.34	3.35	3.33	3.49
Valley estimate with 95 % conf. interval	-	2.97	2.84	2.97

VIII. CONCLUSIONS

This paper has presented several linear Gaussian model-based load profiling techniques that compactly capture multiple behaviors exhibited by residential customers who have traditionally been assumed to be homogenous. The combination of the modeling strategy and the smart meter advance data has permitted a representation that expresses not only load magnitudes at given times of day but also their variability and how these variabilities influence other times of use. The mixture model framework in which this is embedded allows multiple behaviors to be assumed with the statistically most likely one being used to categorize a given residential customer on a given day. In this way, dynamic customer behavior changes can be captured as they evolve with season or changes in routine. Such models have theoretical properties that permit ready use of sampling techniques that have been used to demonstrate gains in accuracy over existing load profile techniques. Such improvements are essential in the

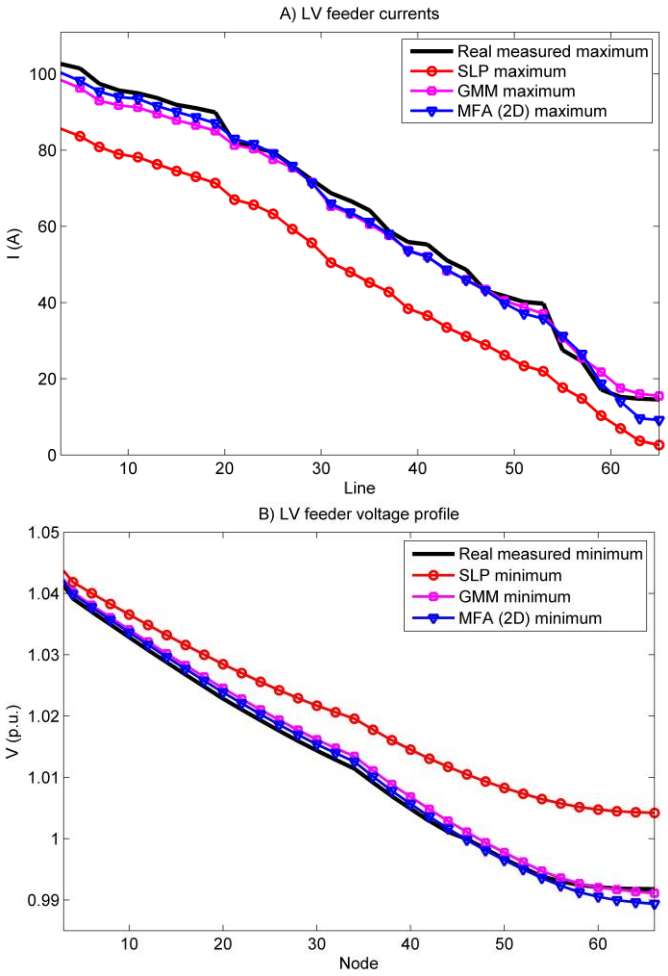


Figure 8: Load flow comparison between SLP, GMM and MFA models.
A) Maximum currents on the LV main feeder (phase L1)
B) Minimum voltages on the LV main feeder (phase L1)

management of smaller and islanded power systems. Loss of performance in the MFA model may have stemmed from overfitting the covariance matrices. In further work, this could be prevented by considering a Bayesian formulation of MFA such as that proposed by [31], which has been shown to provide a more reliable estimate of optimal subspace dimensions. Attention should also now be turned to employing the computationally tractable Gaussian models in temporal and spatial models that could augment emerging state estimation tools [32] and models of regional energy density [33]. Both applications are increasingly important on LV networks as emerging services, such as storage, distributed generation, and demand response measures reach ever-higher penetration levels.

REFERENCES

[1] J. Partanen, P. Juuti, and E. Lakervi, "A PC-based network information system for power distribution companies and consultants," presented at the Int. Conf. High Technol. Power Ind., Tainan, Taiwan, Mar. 1991.

[2] R. Yao & K. Steemers, K. *A method of formulating energy load profile for domestic buildings in the UK. Energy and Buildings*. New York: Elsevier, 2005, vol. 37, pp. 663–671

[3] B. Alexander, "Smart meters, real time pricing, and demand response programs: implications for low income electric customers", Oak Ridge National Lab, Oak Ridge, TN, USA, Tech. Rep., 2007.

[4] Load Profiles and their use in Electricity Settlement. 2013. [Online]. Available: http://www.elexon.co.uk/documents/participating_in_the_market/market_guidance_-_industry_helpdesk_faqs/load_profiles.pdf

[5] A. Seppälä, "Load research and load estimation in electricity distribution," Ph.D. dissertation, Helsinki Univ. Technol., Espoo, Finland, 1996.

[6] "Domestic electricity consumption 2006," Research report (in Finnish) Adato Energia Ltd, 2008.

[7] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 548–555, May 2005.

[8] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification", *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.

[9] A. Mutanen, M. Ruska, S. Repo, and P. Järventausta, "Customer classification and load profiling method for distribution systems," *IEEE Trans. Power Del*., vol. 26, no. 3, pp.1755–1763, July 2011.

[10] B. Stephen and S. Galloway, "Domestic load characterization through smart meter advance stratification," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1571–1572, Sep. 2012.

[11] ESMA, "2009 Annual Report on the Progress in Smart Metering. European Smart Metering Alliance (ESMA) report," Jan. 2010.

[12] REG 1.3.2009/66. Valtioneuvoston asetus sähköntoimitusten selvityksestä ja mittauksesta (Finnish council of state act on electricity settlement and metering).

[13] P. Järventausta, S. Repo, A. Rautiainen, and J. Partanen, "Smart grid power system control in distributed generation environment," *Elsevier Annu. Rev. Control*, vol. 34, pp. 277–286, 2010.

[14] Dept. Energy Climate Change (DECC), Crown Office, "Smart metering implementation programme: Response to prospectus consultation. Overview Document," 2011 [Online]. Available: https://www.ofgem.gov.uk/publications-and-updates/smart-metering-%E2%80%93-response-prospectus-consultation

[15] K. Gram-Hanssen, "Standby consumption in households analysed with a practice theory approach," *J. Ind. Ecol.*, vol. 14, no. 1, 2009.

[16] A. Druckman and T. Jackson, *Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model Energy Policy*. London, U.K.: Elsevier, 2008, pp. 3177–3192.

[17] J. Stephenson, B. Barton, G. Carrington, D. Gnoth, R. Lawson, and P. Thorsnes, *Energy cultures: A framework for understanding energy behaviours. To Appear, Energy Policy*. New York: Elsevier.

[18] V. Neimane, "Distribution network planning based on statistical load modeling applying genetic algorithms and Monte-Carlo simulations," presented at the IEEE Power Tech Conf., Porto, Portugal, Sep. 2001.

[19] S.W. Heunis and R. Herman, "A probabilistic model for residential consumer loads," *IEEE Trans. Power Syst.*, vol. 17, no. 3, Aug. 2002.

[20] E. Carpaneto and G. Chicco, "Probabilistic characterisation of the aggregated residential load patterns," *IET Gen., Transm. Distrib.*, 2007.

[21] R.E. Kass and A.E. Rafferty, "Bayes factors," *J. Amer. Stat. Assoc., Amer. Stat. Assoc.*, vol. 90, no. 430, pp. 773–795, Jun. 1995.

[22] I.T. Jolliffe, *Principal Components Analysis*, ser. Springer Statistics, 2nd ed. New York, USA: Springer, 2002.

[23] B.S. Everitt, *Introduction to Latent Class Models*. New York: Chapman & Hall, 1984.

[24] Z. Ghahramani and G.E. Hinton, "The EM algorithm for mixtures of factor analyzers," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-96-1, 1996.

[25] A.P. Dempster, N.M. Laird, and D.B. Rubin, *J. Royal Stat. Soc. (Method.)*, ser. B, vol. 39, no.1, pp. 1–38, 1977.

[26] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, 1999.

[27] M. Friendly and E. Kwan, *Effect Ordering for Data Displays. Computational Statistics and Data Analysis*. New York: Elsevier, 2003, vol. 43, pp. 509–539.

[28] E. Lakervi and E.J. Holmes, *Electricity Distribution Network Design*, 2nd ed. London, U.K.: Peregrinus, 1995, 325 p.

[29] S. Ingram, S. Probert, and K. Jackson, The impact of small scale embedded generation on the operating parameters of distribution networks. PB Power, Oct. 2003. [Online]. Available http://webarchive.nationalarchives.gov.uk/20100919181607/http:/www.ensg.gov.uk/index.php?article=99

[30] N. Thomopoulos and A. Johnson, "Tables and characteristics of the standardized lognormal distribution," in *Proc. Annu. Meeting Dec. Sci. Inst.*, 2003, pp 2379–2384.

[31] Z. Ghahramani and M.J. Beal, "Variational inference for Bayesian mixtures of factor analysers," *Neural Inf. Process. Syst.*, vol. 12, 1999.

[32] R. Singh, B.C. Pal, and R.A. Jabr, "Statistical representation of distribution system loads using Gaussian mixture model," *IEEE Trans. Power Syst*., vol. 25, no. 1, pp. 29–37, Feb. 2010.

[33] Z. Vale, H. Morais, and N. Pereira, "Energy resources scheduling in competitive environment," presented at the CIRED, Frankfurt, Germany, Jun. 6–9, 2011.

Bruce Stephen (M '09) received the B.Sc. degree in aeronautical engineering from Glasgow University, Glasgow, U.K., and the M.Sc. degree in computer science and the Ph.D. degree in information retrieval from the University of Strathclyde, Glasgow.

Currently, he is a Senior Research Fellow within the Institute for Energy and Environment, University of Strathclyde, Glasgow, U.K. and is a Chartered Engineer. His research interests include distributed information systems and machine learning applications in power system condition monitoring and asset management.

Antti Mutanen was born in Tampere, Finland, on June 10, 1982. He received his M.Sc. degree in electrical engineering from Tampere University of Technology, Tampere, Finland, in 2008.

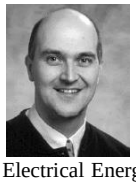
Currently, he is a Researcher and a Postgraduate Student in the Department of Electrical Energy Engineering, Tampere University of Technology. His main research interests are load research and distribution network state estimation.

Stuart Galloway received the M.Sc. and Ph.D. degrees in mathematics from the University of Edinburgh, Edinburg, U.K., in 1994 and 1998, respectively.

Currently, he is a Senior Lecturer within the Institute for Energy and Environment, University of Strathclyde, Glasgow, U.K. His research interests include power system optimization, numerical methods, and simulation of novel electrical architectures.

Graeme Burt (M'09) received the B.Eng. degree in electrical and electronic engineering from the University of Strathclyde, Glasgow, U.K., in 1988 and the Ph.D. degree in fault diagnostics in power system networks from the University of Strathclyde in 1992.

Currently, he is a Professor at the University of Strathclyde, and is Director of the University Technology Centre in Electrical Power Systems, sponsored by Rolls Royce.



Pertti Järventausta (M'12) received the M.Sc. degree in electrical engineering from Tampere University of Technology, Tampere, Finland, in 1990 and the Dr.Tech. degree in electrical engineering from Lappeenranta University of Technology, Lappeenranta, Finland, in 1995.

Currently, he is a Professor at the Department of Electrical Energy Engineering, Tampere University of Technology. His main interest focuses on electricity distribution and electricity market.

Corrigendum, 7.9.2017

in respect of [P7], Section VI-B, reference to Fig. 6.

The reference to Fig. 6 in Section VI-B refers to a missing figure, not to the Fig. 6 located on page six. The reference to this missing figure was left to the paper accidentally after the figure had been removed to shorten the paper. The removed figure is shown below.

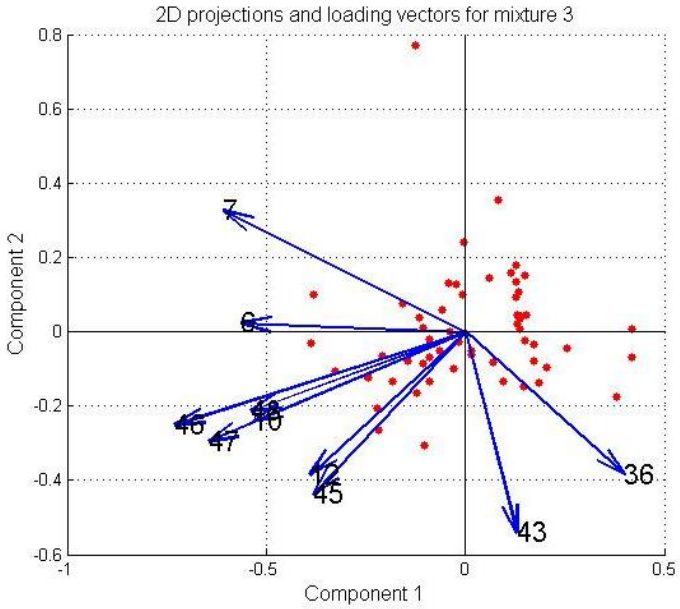


Figure 6: Vector plot representation of a factor analyzer loading matrix.

Later in Section VII-A, the Fig. 6 located on page six is referred to correctly.

Publication 8

P. Koponen, A. Mutanen, and H. Niska, "Assessment of some methods for short-term load forecasting," presented at the 5th IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Istanbul, Turkey, Oct. 12–15, 2014.

Available at: https://doi.org/10.1109/ISGTEurope.2014.7028901

Assessment of Some Methods for Short-Term Load Forecasting

Pekka Koponen, Member IEEE
VTT Technical Research
Centre of Finland
Espoo, Finland
pekka.koponen@vtt.fi

Antti Mutanen
Tampere University of Technology
Tampere, Finland
antti.mutanen@tut.fi

Harri Niska
University of Eastern Finland
Kuopio, Finland
harri.niska@uef.fi

***Abstract*— Accurate forecasting of loads is essential for smart grids and energy markets. This paper compares the performance of the following models in short-term load forecasting: 1) smart metering data based profile models, 2) a neural network (NN) model, and 3) a Kalman-filter based predictor with input nonlinearities and a physically based main structure. The comparison helps method selection for the development of hybrid models for forecasting the load control responses. According to the results all these three modeling approaches show much better performance than 4) the traditional load profiles and 5) a static outdoor temperature dependency model applied with a lag. The neural network model was the most accurate in the comparison, but the differences of the three methods developed were rather small and also other aspects and other methods must be considered and compared when selecting the method for a specific purpose.**

***Index Terms*-- power demand, demand forecasting, load modeling, prediction algorithms, artificial neural networks.**

I. Introduction

Accurate estimation and forecasting of loads is a necessary enabler for the development of smart grids, energy markets and customer engagement. Starting from 2014 hourly interval metered consumption data of almost every consumer are recorded in Finland. With it the accuracy of load models and forecasts can be much improved and different methods are being developed for the purpose.

In the literature there are many papers that describe various approaches to short-term load forecasting. Load forecasting methods are reviewed in [1] except physically based load response models that are initially reviewed in [2]. There are also papers on merging different approaches, e.g. [3,4]. In 1989 a comparison of five short-term load forecasting methods was published [5]. It included a state space method with Kalman-filter, but without any physically based structure or nonlinearities.

In this paper the focus is on short-term forecasting of the total power of a large group of residential customers. It compares the performance of three different approaches: 1) load profiling method based on smart metering data and clustering, 2) partly physically based model comprising a physically based main structure and a Kalman-filter based predictor with input nonlinearities, and 3) a Multi-Layer Perceptron (MLP) neural network. For comparison, standard customer class load profiles and a static polynomial fit with a lag were also included in this study, because some similar load forecasting approaches are still applied in the industry. Results are summarized with tables using performance indices Sum of Squared Errors (SSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), etc.

II. Definition of the Problem

A. The Forecasting Task

The prediction task studied was at 9 a.m. to predict as accurately as possible the next day total consumption of a large group of small houses and apartments. It includes forecasting both the hourly power and the daily energy.

B. Objectives of the Method Comparison

The objectives of the comparison included

- screening of methods for further development,
- learning about the relative merits and improvement potential of the methods, and
- comparing the performance with some approaches known to be applied in the industry.

C. The Data

The data set used in this study comprised two years (2009-2010) of

- hourly interval measured consumption of 3516 individual small customers i.e. electricity consumers,
- measured outdoor temperature representative to the studied power distribution area, and
- local outdoor temperature forecast available each day at 9 a.m. for each hour in the load forecasting horizon.

Only consumers with measured hourly power always under 50kW were included in the data set. The data of the year 2009 were used in load model identification and the year 2010 data were reserved exclusively for verification. The data are shown

This research was financially supported by the Finnish Smart Grids and Energy Markets (SGEM) research program 2009-2014 of the Cluster for Energy and Environment (CLEEN). The measurement data were provided by the distribution operator KSS (Koillis-Satakunnan Sähkö Oy).

in Figures 1 and 2. In the temperature measurement of 2010 there are two gaps. In model verification those gaps were filled with the latest temperature forecasts. In the model identification and verification data the temperature ranges were from -24.7 to 27.6 degrees C and from -29.5 to 32.8 degrees C respectively. Thus we can compare the performance of the models also outside the identification range.

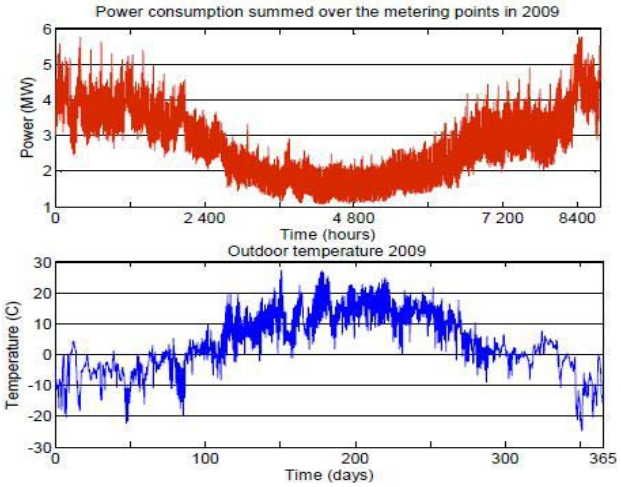


Figure 1. The data used for model identification (2009).

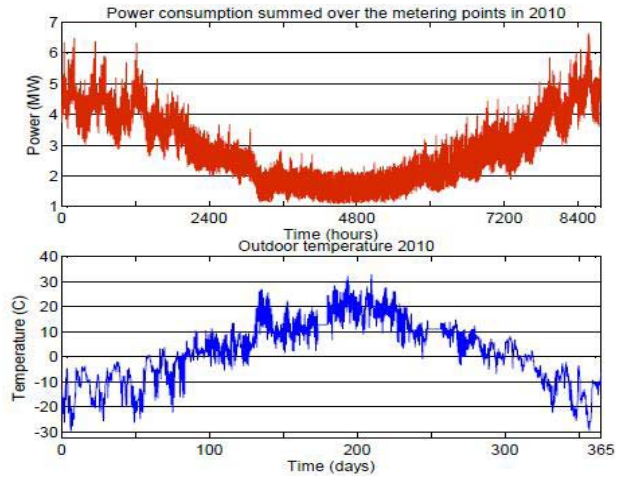


Figure 2. The data used for model verification (2010).

III. The Methods Compared

The following approaches or load models were compared.

A. Load Profiles

Customer class load profiles are widely used in distribution network analysis, and in electricity retail risk management and sales. The purpose of customer class load profiles is to give estimates for the customer level loads and their variability which can then be aggregated to higher level estimates. In network planning and operation, the load profiles are often used to predict future network loadings

1) Standard Customer Class Load Profiles

Finnish Electricity Association Sener (which later merged to Finnish Energy Industries) has defined standard customer class load profiles for 46 different customer groups. Their usage is described in detail in [6].

The customer classification is usually stored in the electric utility customer information system (CIS). In this case, the studied customer group contained customers from 25 different customer groups. A sum load profile was calculated for the studied group of customers using these 25 standard customer class load profiles and annual energies measured during the identification year.

The advantage of using standard customer class load profiles is that all the necessary information needed for calculating load forecasts is already available in the existing distribution management systems. The downside is that the forecasting accuracy is hindered by the following factors:

- The electricity consumption habits are constantly changing but the load profiles and customer classification are rarely updated.
- The load profiles lack models for responses to outdoor temperature and dynamic load control actions that are applied based on market or network.
- Exceptionally behaving customers cannot be modeled with the standard customer class load profiles.
- Geographical differences in electricity consumption are not modeled since the standard customer class load profiles are used nationwide.

2) Cluster Load Profiles

In order to address shortcomings of the standard profiles, we have developed a load profiling method that utilizes hourly consumption measurements from previous year(s) and updates both the customer class load profiles and customer classifications [7]. The customers are grouped into similarly behaving groups with the help of a K-means clustering algorithm and cluster load profiles are calculated for each cluster. In this paper, the original customer classification was used as a starting point for the clustering and exceptionally behaving customers were separated for individual load profiling. This resulted in 25 cluster load profiles and 34 individual load profiles. Similarly to the standard customer class load profiles, these profiles contained expectation and standard deviation values for each hour of the year.

Seasonal temperature dependency parameters (%/°C) were calculated for each cluster using the method presented in [8]. Temperature dependency parameters together with measured (when applicable) and forecasted outdoor temperatures from 24 previous hours were used to adjust the hourly load forecasts.

When combined with temperature measurements and forecasts, the smart metering based cluster load profiles provide much better forecasts than the standard customer class load profiles. The forecasting accuracy of the cluster profiles still falls behind the best online forecasting methods but it has other beneficial properties. The cluster load profile method does not require continuous access to smart meter data. Delays or interruptions in smart meter reading do not matter and the measurement database needs to be accessed only once a year

when the cluster load profiles are updated. Only the outdoor temperature measurements and forecasts need to be available when the next day forecasts are made. The cluster load profiles can also be used in the existing network calculation software with very little changes.

The sum load could also be modelled with a separate sum load profile similar to the individual load profiles. However, in this case the aggregated cluster profiles performed better. MAPE for the sum load profile was 4.38 % while the MAPE for the aggregated cluster profiles was 4.09 %.

B. Neural Network Model

The advantage of neural network (NN) models compared to other statistical methods is that they are able to learn complex, nonlinear, and a priori unknown relationships between input and output variables from the training data [9]. On the other hand, NN models are often difficult to interpret, highly complex and not transparent.

Regarding the NN model we used the feed-forward Multi-Layer Perceptron (MLP) network. The choice was based on its simplicity and accuracy shown in previous studies [1]. MLP consists of a network of simple processing elements (neurons) and connections. Neurons are arranged in layers, namely the input layer, the hidden layers, and the output layer. Each neuron computes a weighted sum of the inputs, processes this using a neuron transfer function (called also as activation function; it should not be confused with a transfer function for a linear dynamical system) and distributes the result to the subsequent layer. The output signal of a single neuron can be expressed as:

$$y = f\left(\sum_{i=1}^{n}(w_{ij}x_i + b_j)\right) . \quad (1)$$

Where f denotes the neuron transfer function, j is the index of the neuron, n is the number of neurons in input layer, x_i is the input from i^{th} input neuron, w_{ij} is the weight between i^{th} input neuron and j^{th} hidden neuron and b_j is the bias of the neuron.

Training of the MLP network is performed using the Back-Propagation (BP) algorithm, which adjusts iteratively the weights of the network to minimize the error function, namely the squared errors calculated between actual and desired outputs. Regularization, such as so called early stopping, is adopted to control over-fitting.

In this study, the standard MLP network with one hidden layer was employed to learn a functional (non-linear) relationship between input and output variables in order to perform predictions. In the model set-up, the output variable consisted of the hourly power at time to be forecasted. The input variables comprised well-known predictor variables, i.e. timing variables (day of year, day of week, hour of day; all of them transformed into continuous form) at time to be forecasted, the length of day at time to be forecasted, as well as lagged ambient temperature values (either forecasted or measured) available at time when forecasting occurs i.e. at 9 a.m. previous day. Time-lags of ambient temperature were determined empirically between 5 and 40 hours, finally with 5 hours interval. In general, the selected input variables aim at describing temporal rhythm, light and temperature dependence, as well as temperature delay of hourly electric loads of a customer group.

The proposed MLP network model was trained using Levenberg-Marquardt (LM) algorithm through 3000 training epochs. For controlling the over-fitting, the standard early-stopping strategy was adopted by stopping the training when the internal error of the network calculated from the identification/training data increased for 25 iterations. The selection of feasible architecture of the network was based on experimental tests, which showed that one hidden layer with 15 hidden nodes, sigmoid transfer functions for hidden units, and linear transfer function for output are sufficient.

C. Kalman-filter Based Predictor with Input Nonlinearities

For this method comparison, the physically based model main structure approach described in [10] was applied to the data of this comparison. The model was built for the aggregated sum of all the customers. Possible improvement of the forecasting accuracy by clustering was not yet studied. The structuring of the model into parallel linear model components and their input nonlinearities was designed based on physical information of the main load types. The submodels included in the main structure are:

- electrical heating (transfer function model)
- electrical cooling (transfer function model)
- day length dependent lighting
- constant load component (constant)
- weekly rhythm for the year (a week long time series).

Adding some other submodels, such as an air-to-air heat pump model, were also tested, but abandoned in this case, because they did not improve the identification accuracy adequately.

Two of the included submodels (heating and cooling) have outdoor temperature as an input variable and consist of input saturation and a transfer function for a linear dynamical system. The model structure allows adding more complex static monotonic input nonlinearities that may still improve the performance. Now this possibility was not even tried, because it was considered better to keep model identification and comparison as simple and clear as possible.

The linear dynamics are described by transfer functions of the form

$$G_i(s) = y_i(s)/u_i(s) . \quad (2)$$

Where $s = j\omega$ and $u_i(s)$ and $y_i(s)$ are polynomials of s for the input and output respectively for the submodel i. Each submodel i has also a static input nonlinearity defined by function $f_i(u)$, where u is the input signal to the whole model,

$$u_i(t) = f_i(u) . \quad (3)$$

The model output y is the sum of the submodel outputs y_i

$$y(i) = \sum_{i=1}^{N} y_i(t) \quad , \qquad (4)$$

where N is the number of submodels.

The submodels were identified one by one and the output of the earlier identified submodels was subtracted from the output y(t). Minimizing SSE was the objective of the identification. The transfer functions were converted and combined to state space form and a Kalman-filter based predictor [10, 11] was designed using the Matlab® function *kalmd*. The covariance matrices for the process noise and the measurement noise were identified based on the identification data and not updated during the verification phase. Thus constant Kalman gain is applied during the forecasting, which improves the robustness but may also reduce the accuracy.

D. Static polynomial fit with a lag

Static polynomial of order 4 and best lag was fitted and applied for the yearly load dependency on outdoor temperature. 8 hours lag gave the best fit. Submodels for annually identified weekly rhythm and day length dependent lighting load were included. Using seasonal weekly rhythm models did not improve forecasting accuracy of the static polynomial fit method nor the Kalman-filter based predictor.

In extreme temperatures the polynomial causes big errors with the verification data. Applying temperature saturation to the curve slightly improves the forecasting performance and robustness, but it is not obvious how the saturation limits should be defined. Thus we did not apply saturation limits in the comparison.

E. Summary of the methods compared

The methods included in the comparison are recapped and abbreviated in Table I. It was found out that the weather forecast improves the performance of all the methods substantially. For example, for the partly physically based method it improved the MAPE of hourly power forecasts from 7.37 % to 4.57 %. Thus only the results with the weather forecasts are shown in the following. Possible reporting and discussing the impacts of the accuracy of weather forecasts on load forecasting performance is left to a future study.

TABLE I. THE METHODS COMPARED

Method	Short Name
Standard customer class load profiles (by SENER)	SENER load profiles
Best lag (8h lag) static temperature dependency fit with 4th order polynomial.	Static polynomial & lag
Cluster load profiles with seasonal temperature dependency for daily energy (Collection of linear models)	Cluster load profiles
Kalman-filter predictor in a physically based component model structure (linear submodel dynamics with input saturations)	Partly physically based
Neural network model with time-lagged temperature values	Neural network model

IV. COMPARISON OF PERFORMANCE

A. Verification of performance indices

The external forecasting performance of the methods was measured by comparing the forecasts and actual measurements during the verification year. The forecasts were normalized so that value 1 represents the time average of the observed total load of the test group. Then the following performance indices were calculated:

- Sum of squared error of prediction (SSE) also known as the sum of squared residuals
- Root Mean Square Error (RMSE) = root(mean(e_t^2))
- Mean Absolute Error (MAE) = mean ($|e_t|$)
- Mean Absolute Percentage Error (MAPE) that is the mean of absolute errors divided by the observed values (= mean($|p_t|$), where $p_t = 100\ e_t / y_t$,where y_t is the observation at time t).

Here e_t is the forecasting error at time t. For more information on measures of forecast accuracy read [12].

B. The results of the comparison

Table II compares the verification results of the methods in forecasting hourly energy and daily energy. The performance index value differences between the cluster load profiles and the partly physically based model do not exceed 90% confidence interval estimates. All the others do.

TABLE II. PERFORMANCE IN FORECASTING

	daily energy forecast, normalised				
Short name	***SSE***	***RMSE***	***MAE***	***MAPE %***	***Std***
SENER load profiles	7.957	0.1477	0.1046	10.07	0.137
Static polynomial & lag	2.016	0.0743	0.0568	5.70	0.076
Cluster load profiles 1	0.874	0.0489	0.0335	3.09	0.047
Partly physically based	0.804	0.0469	0.0322	3.07	0.047
Neural network model	0.619	0.0412	0.0277	2.59	0.041

	hourly energy forecast, normalised				*annual energy error %*
Short name	***SSE***	***RMSE***	***MAE***	***MAPE %***	
SENER load profiles	275.8	0.1774	0.1345	0.00	-5.45
Static polynomial & lag	89.8	0.1012	0.0807	0.00	-0.53
Cluster load profiles	33.6	0.0620	0.0428	0.00	-1.19
Partly physically based	34.0	0.0623	0.0454	0.00	-0.07
Neural network model	27.8	0.0564	0.0393	0.00	-0.64

C. Time series of forecasting errors

Figure 3 compares the time behavior of the forecasting errors of the methods (error = forecast - measured). Due to the limited resolution of the figures the envelope is seen rather than the individual values of the forecasts. In summer the relative differences in forecasting performance are clearly visible, while in the middle of winter the neural network does not seem to perform much better than the other methods.

D. Errors as a function of power

It was also analyzed how the size of errors depends on the power at the same moment. For all the methods the errors

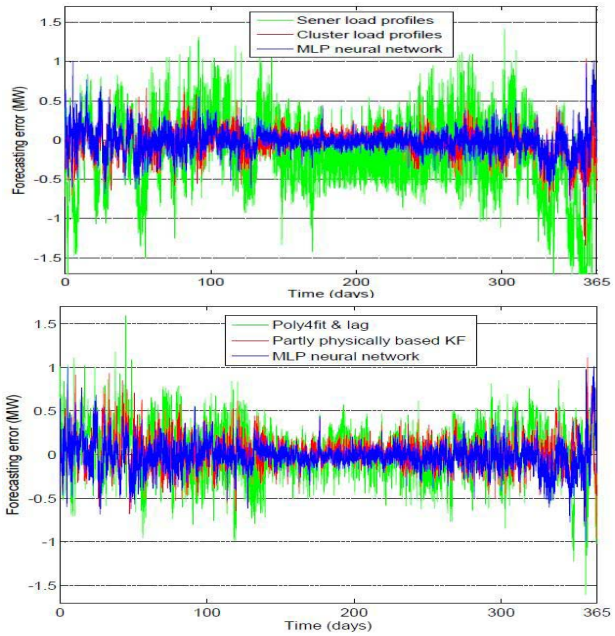


Figure 3. Time series behavior of forecasting errors of the five methods.

grow as the power increases and the mutual differences in this behavior are small and maybe insignificant, see Figure 4. The cluster load profile method has some large negative errors during high power situations but otherwise its errors roughly equal the errors of the neural network method. See also Figure 5. The partly physically based method had slightly smaller errors than the other methods in high power situations but in all other situations slightly bigger errors.

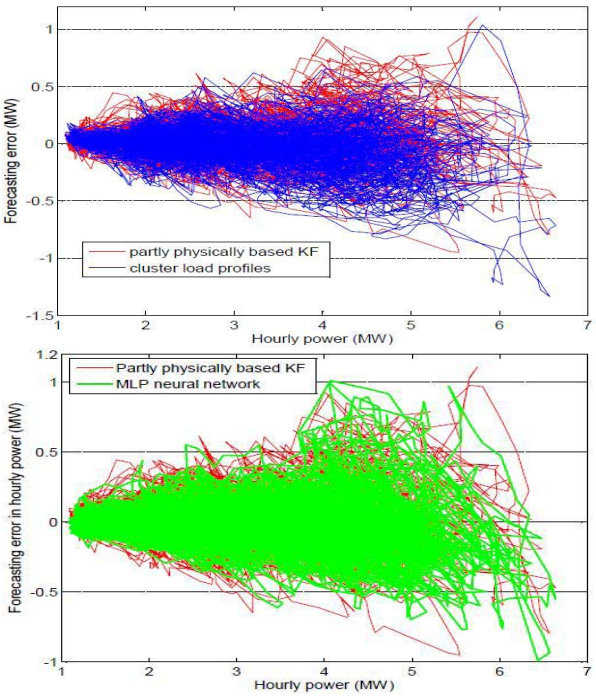


Figure 4. The dependence of forecasting errors on power.

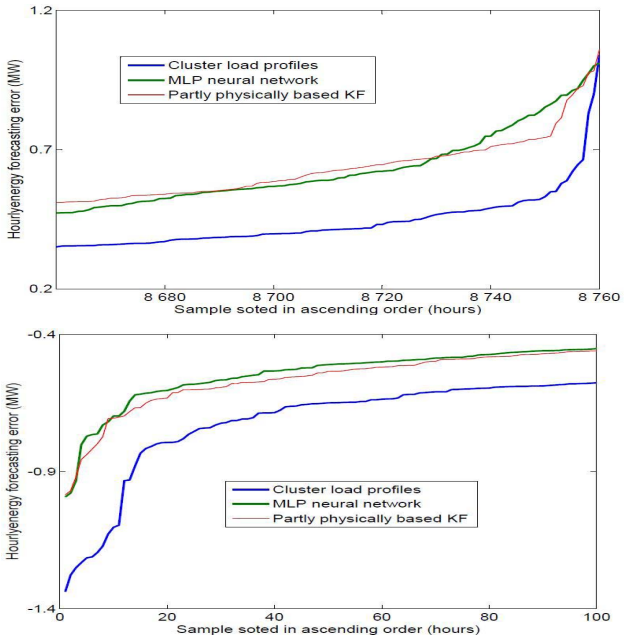


Figure 5. The biggest negative and positive forecasting errors, when the x axis is sample hours ordered.

Figure 6 shows time series of the biggest forecasting errors for 1) partly physically based, 2) cluster load profiles, and 3) neural network methods respectively. It also gives the load to be forecasted and the ambient temperature. All the models have their biggest errors during the same special days at Christmas time, when it was also cold.

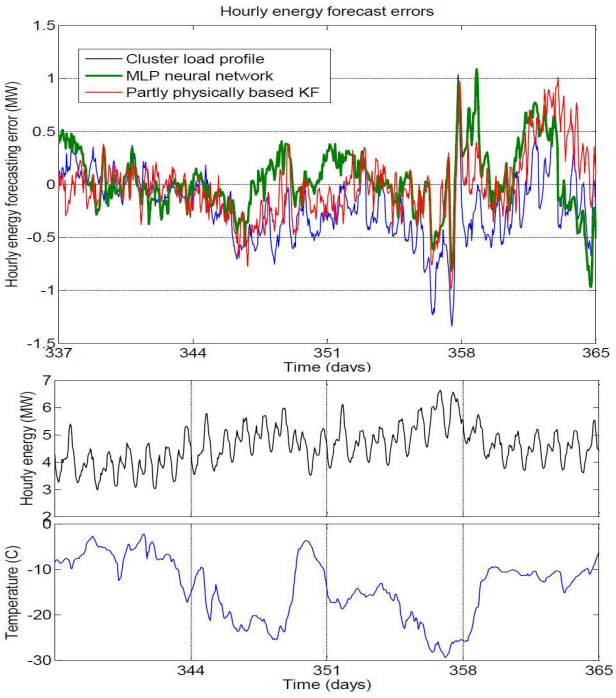


Figure 6. The biggest forecasting errors for 2010 are on 22-24 December when ambient temperature is low and the load high.

E. Assessment of the comparison and further steps

It may is useful to develop new performance indices that better take into account the real needs of the specific real situation. For example, accurate forecasts are needed most when the network loads or electricity whole sale market prices are high.

Because cheaper night time tariffs and partially storing electric night time heating are applied in the studied area, the daily load peak is near midnight, but the highest whole sale market area prices were outside the night tariff.

In this comparison the NN model was the most accurate especially outside the peak load times. The cluster load profile was second in accuracy in forecasting hourly powers, but in forecasting the daily energies the physically based model was slightly better regarding most of the performance indices. During the critical times in winter the differences in accuracy between the three main methods of the comparison were small except for a small number of hours where the clustering load profiles method forecasted too small loads.

The load profiling approaches do not require continuous measurement of power. Adding feedback from power measurement removed the annual energy error and improved the other indices, such as the hourly energy forecast SSE from 33.6 to 29.0. The other methods assume that real time measurement of the sum power of the target load group is available at the forecasting moment. Starting at the beginning of the year 2014 in Finland the hourly consumption of each customer is read every night and after some hours the previous day measurements are available for load forecasting. Combining those measurements with real time measurements from the distribution network enables estimation of the aggregated power of the target group. The dependence on the reliability and accuracy of the real time information is an important issue to consider and to study further.

It is increasingly important that the models predict the control responses. Physically based load forecasting models can do that [13]. It is still unclear how and to what extent they can be added to artificial neural network model or load profiles. Forecasting control responses is a relevant topic for future research. Physically based models include a-priori information on the system which can help to maintain good forecasting performance also in situations not included in the identification data set. In this study the temperature range in verification was wider than in the identification and there the accuracy of the partly physically based model was only slightly and not significantly better. The results of the comparison also indicate a need to study how the physically based model can be improved regarding the seasonal variations in the daily load profile.

In our initial studies with data from some other network areas and years, it seems that the methods work well also with them, but the relative order regarding the accuracy indices may vary slightly. Further data and research are needed to confirm that.

The predictions by the neural network model were the most accurate in terms of the performance indices of Table II. New situations that have not been experienced before may include uncertainty, challenges and risks regarding the prediction performance of purely data-driven NN models. In addition, instead of the conventional MLP network it is necessary to test more advanced NN and machine learning techniques such as support vector machines and hybrid methods as well. As a reference for the comparison standard regression models combined with input nonlinearity are needed, too. These and analysis of confidence intervals are part of the planned future studies.

In order to better learn the benefits and limitations of different approaches extensive evaluations are still needed using data from several years and distribution areas. The methods compared and their tuning and evaluation should also be harmonized. Based on the literature, such as [1,3,4,13], there are ample improvement possibilities to study.

V. CONCLUSION

Some promising methods were compared in short-term load forecasting. The neural network was the most accurate in this comparison, but the differences in performance were rather small and the other methods have their inherent strengths. We plan to study, merge and evaluate all the three main approaches further.

REFERENCES

[1] H. Hahn, S. Meyer-Nieberg, and S. Pickl, “Electric load forecasting methods: tools for decision making,” European Journal of Operational Research, Vol. 199, pp. 902–907, 2009.

[2] P. Koponen, Measurements and models of electricity demand responses, Research report VTT-R-09198-11, VTT Espoo, 24 p., 2012. http://www.vtt.fi/inf/julkaisut/muut/2012/VTT-R-09198-11.pdf

[3] J. M. C. Sousa, L.M. P. Neves, H.M.M. Jorge, "Short-term Load Forecasting using information obtained from Low Voltage Load Profiles," POWERENG 2009, Lisbon Portugal, March 2009.

[4] A. Karsaz, H. Khaloozadeh, "Medium Term Horizon Market Clearing Price and Load Forecasting With an Improved Dual Unscented Kalman Filter," IEEE International Conference on Control and Automation, Guangzhou, CHINA - May 30 to June 1, 2007, pp. 507--513.

[5] I. Moghram, S. Rahman, "Analysis and Evaluation of Five Short-Term Load Forecasting Techniques," IEEE Transactions on Power Systems, Vol. 4, No. 4, pp. 1484-1491, October 1989.

[6] A. Seppälä, Load research and load estimation in electricity distribution, Ph.D. dissertation, Helsinki University of Technology. 1996.

[7] A. Mutanen, P. Järventausta, M. Kärenlampi and P. Juuti, ”Improving Distribution Network Analysis with New AMR-Based Load Profiles,” 22nd International Conference on Electricity Distribution, CIRED, Stockholm, June 2013.

[8] A. Mutanen, M. Ruska, S. Repo and P. Järventausta, 2011, "Customer Classification and Load Profiling Method for Distribution Systems,” IEEE Transactions on Power Delivery, Vol. 26, No. 3.

[9] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd Edition. Prentice-Hall, Upper Saddle River, NJ, 1999.

[10] P. Koponen, "Short-Term Load Forecasting Model Based on Smart Metering Data," IEEE SG-TEP 2012, Nuremberg, December 2012.

[11] P. S. Mayback, Stochastic models, estimation, and control, Vol. 1, Academic Press, 1979.

[12] R. J. Hyndman, A.B., Koehler, "Another look at measures of forecast accuracy," Monash University, 2005. . http://www.robjhyndman.com/papers/mase.pdf

[13] P. Koponen , J. Saarenpää (eds.), Load and response modelling workshop in project SGEM. 10 November 2011, Kuopio. Espoo, VTT, VTT Working Papers 188, 2011. . http://www.vtt.fi/inf/pdf/workingpapers/2011/W188.pdf

Publication 9*

A. Mutanen, P. Järventausta, and S. Repo, “AMR-based load profiles and their effect on distribution system state estimation accuracy,” submitted to International Review of Electrical Engineering.

* Not included in the web version of the dissertation due to pending review.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-4094-3
ISSN 1459-2045