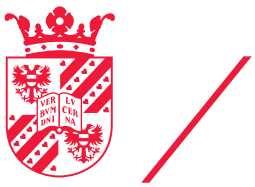
Enhanced Observability of MV Distribution Networks using Personalized Differential Private Smart Metering



university of groningen

faculty of mathematics and natural sciences

Rutger Prins

First Supervisor: Dr. Pietro Tesi

Second Supervisor: Prof. Dr. Claudio de Persis

Faculty of Science and Engineering
University of Groningen

This dissertation is submitted for the degree of
Master of Science

Research Project July 2017

Abstract

State estimation is essential for the monitoring and control of power grids. The observability of a 33-bus medium-voltage power grid is enhanced using measurements of smart meters at dwellings. where personalized differential privacy (PDP) is guaranteed using the Sample Mechanism in conjunction with the Laplace Mechanism. The Sample Mechanism samples measurements with probability dependent on the individual privacy budget, which is seen as intermittent transmission of measurements. State estimation is performed using the Extended Kalman Filter and missing measurements are imputed with the mean of the measurements at a bus. A power load generator for Irish dwellings is used to create synthetic datasets. Three scenarios are considered, where all dwellings trust the aggregator fully, partly or not at all to guarantee their differential privacy. For the untrusted scenario, the sample mechanism is effective and increases the state estimation accuracy. For the partly and fully trusted scenario, the state estimation accuracy decreases and the Sample Mechanism should be avoided in order to achieve higher state estimation accuracy. Also, the effect of varying privacy aspects and dataset characteristics are researched and directions for further research are given.

Table of contents

Chapter 1

Introduction

In this chapter, an introduction of concepts in this thesis, motivation, relevant work and the focus of this research are introduced. Also, a short background on the subject and the outline of this thesis is given.

1.1 Introduction

State Estimation (SE) is an essential component of electric power grid monitoring and control systems [61, 52, 28]. Due to the limited metering infrastructure, low-voltage distribution grids have limited observability [9]. Currently, distribution grids are observed using load, voltage magnitude and voltage angle measurements at a few buses. Due to volatile distributed renewable generation, the pseudo measurements are not accurate enough for efficient and safe operation of distribution grids [52]. These include voltage conservation and VAR control (VVC) and fault location, isolation and restoration (FLIR). A dedicated real-time monitoring infrastructure to monitor and control distribution grids may be too expensive [52, 33].

Smart meters are world-wide deployed at dwellings for billing purposes and could provide the network operator with measurements to enable state estimation of the distribution grid. Such detailed, up-to-date visibility about the consumption patterns enables important new ways for balancing supply and demand, controlling peak load and improving delivery reliability and efficiency [62]. Smart inverters that can be found in solar panels, energy storage units and electric vehicles can be controlled within milliseconds [9]. The widespread deployment of smart meters have serious privacy implications of residential and industrial customers [40]. A large range of information can be revealed, such as how many people are at home, sleeping routines and eating routines. This needs to be addressed to gain consumer acceptance and trust [22].

Efforts have been made to guarantee differential privacy using smart meters in state estimation of distribution grids [52]. By the contextual integrity theory of Nissenbaum, privacy is provided by appropriate information flows that conform with personal contextual information norms [42]. Appropriateness differs per person and context and therefore differential privacy using personalized privacy budgets is preferred. The idea of not transmitting measurements intendedly to create appropriate information flows in combination with state estimation in distribution grids is not researched yet.

In this thesis, intermittent transmission of measurements as a privacy-preserving method and its effect on state estimation in distribution grids is researched. The privacy-preserving method is combined with the known distributed Laplace noise mechanism. Privacy is quantified using the notion of differential privacy.

1.2 Background

1.2.1 Smart Grid

Smart grid is "an electricity network that can intelligently integrate the actions of all users connected to it - generators, consumers and those that do both - in order to efficiently deliver sustainable, economic and secure electricity supplies" [49]. Domains in the smart grid are [49]: customers, markets, service providers, operators, bulk generation, transmission and distribution. Relations between these domains can be seen in figure 1.1. The smart grid can be seen as an evolution of the current power grid and adds new features to the current power grid [57, 22]:

- **Enabling informed participation by customers**: customers can actively participate in the energy market, as it allows for bi-directional flow of both energy and information. Demand load control based on the power grid conditions is shifting from traditional interruptible demand at industrial plants towards demand-response programs.
- **Accommodating all generation and storage options**: supporting a large number of generation (exploiting natural resources) and storage resources requires anticipation of interruptibility and unavailability, while simultaneously balancing costs, reliability and environmental considerations.
- **Enabling new products, services and markets**: smart grid initiatives receive a lot of investments and enables the development of innovative customer-oriented solutions.
- **Providing the power quality for the range of needs**: the smart grid can provide better power quality for different types of users (e.g. industrial, commercial and

residential customers) by local control of power needs in a micro-grid and supporting distributed generation.

- **Optimizing asset utilization and operating efficiently**: operating efficiency has been dropped by automation and could increase even more by developing robust and reliable maintenance systems that support the complexity of the smart grid.
- **Operating resiliently to disturbances, attacks and natural disasters**: participation of demand-side resource control and distributed generation respond better in case of disturbances and emergencies.

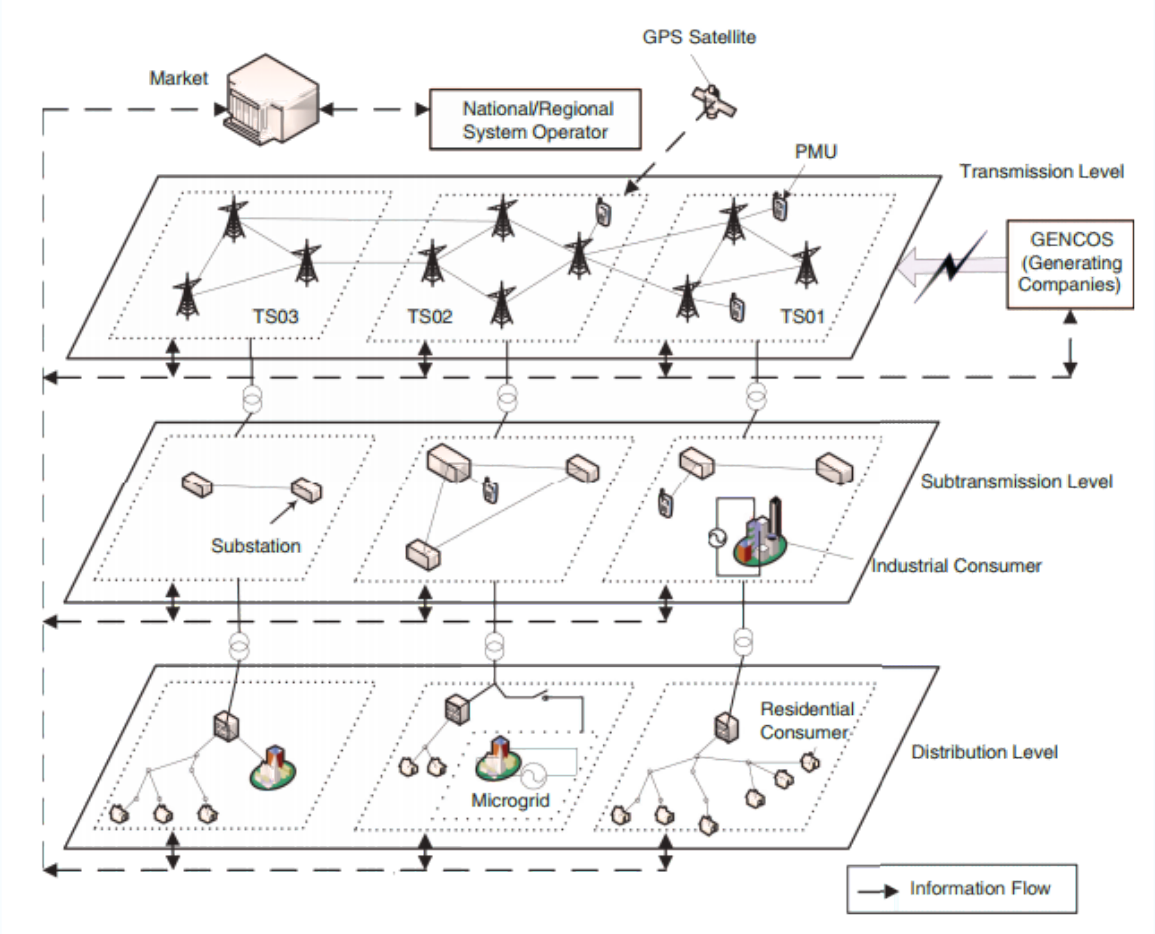


Fig. 1.1 Smart grid domains and its relations [28].

1.2.2 Smart Meters

Smart meters are "advanced meters that identify consumption in more detail than conventional meters and communicate via a network back to the utility for monitoring and billing purposes" and is a component of smart grids [14]. In contrast to advanced meters, smart meters have two-way information flow. Advanced metering infrastructure (AMI) refers to the system of meters and associated communications. It allows demand response: the management of demand in relation of prevailing supply conditions. The most advanced form of AMI is the smart grid, where (remote) load control is applied using very high resolution measurements.

Utilities of smart meters include, but are not limited to [6]: billing optimization, load monitoring and management for specific groups or regions, energy theft/losses detection, load forecasting for specific groups or regions, load forecasting for individual consumers, time-based rates, demand-based rates, individual data analytics, appliance health monitoring [62, 13], approximate employment status of population [4], in-home feedback tools such as estimated bills.

1.3 Outline of Thesis

This thesis starts with an introduction of the research problem and research methodology. In the next chapter, a literature review is given. A mechanism that uses intermittent transmission to provide a personalized differential privacy budget to users is given. Next, in the modeling chapter, the research setup and its validation design are elaborated. Finally, conclusions are drawn and directions for future research are described.

Chapter 2

Preliminary Literature Review

In this chapter, a brief overview of literature on state estimation in distribution grids and differential privacy is given. It is used to frame the research in chapter 3 (Research Methodology).

2.1 State Estimation in Distribution Grids

State estimation in distribution grids can be used to make the medium-voltage network more observable with only measurements at the HV/MV substation and aggregation of individual measurements at the LV side of MV/LV substations [3]. These networks are under-determined as the real-time measurements are insufficient to make the system fully observable. State estimation can be used to find the underlying state of the system with a limited number of measurements. Chosen state variables for enhancing observability are voltage magnitude (denoted as $|V|$) and voltage angle (denoted as V_{angle}) [60].

State estimation accuracy can be measured using mean absolute percentage error (MAPE) and is used by many other papers that elaborate on state estimation in distribution grids [61, 29, 3]. Another measure that can be used is the root-mean-square error (RMSE) [29]. A maximum deviation of 0.7% for the voltage magnitude and 0.7 crad ($0.7 \cdot 10^{-2}$ rad) for the voltage angle is considered acceptable [33], note that this measure is different from MAPE or RMSE.

State Estimation can be performed using static or dynamic filtering. Static filtering does not use any preceding information [58], in contrast to dynamic filtering where consecutive and uncorrelated sets of measurements varying in time are used. Dynamic filtering is more accurate and provides the capability of predicting the next time step [56], which allows operators to take more time for control actions (especially in case of emergencies).

2.2 Differential Privacy

Differential privacy (DP) describes a promise by a data holder or curator [17]: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources are available.". It provides privacy by process by introducing randomness. Randomness is essential as "any non-trivial privacy guarantee that holds regardless of all present or even future sources of auxiliary information". Non-trivial means that there exists a query and two databases that yield different outputs under this query.

Differential privacy has the following desirable qualitative properties [17]:

- **Protection against arbitrary risks**: protects beyond re-identification.
- **Automatic neutralization of linkage attacks**: this includes attacks using auxiliary information from the past, present and future.
- **Quantification of privacy loss**: enables comparison of different mechanisms that provide differential privacy. An example is a comparison of mechanisms versus accuracy.
- **Composition**: enables the quantification of privacy loss over multiple computations.
- **Group privacy**: permits the analysis and control of privacy loss incurred by groups. An example of a group is a family or residents in a care home that could have the same daily routine and thus inferences from one resident can result in privacy loss for another resident.
- **Closure under post-processing**: tells that it is immune to post-processing. A data analyst could not increase privacy loss by using post-processing techniques (e.g. machine learning) or auxiliary information.

Despite the positive qualitative properties, differential privacy does not protect against all harm. It also does not enhance privacy where no privacy exists or guarantee that a secret remain secret in the future when auxiliary information is available for the adversary. Consider a health survey with the goal to discover early indicators of a particular disease and produces conclusive results. It can not be said that differential privacy is violated if an individual that show these indicators and has the disease participated in the survey, as the same correlations will be observed with nearly the same probability with or without the participation of the individual.

2.2.1 Definitions of Differential Privacy

Several definitions of differential privacy have been developed. Four definitions are compared in table 2.1. In this table, definition refers to the parameters used to tune the differential privacy guarantee. Then, the definition is given in terms of ε-differential privacy. Then the differential privacy that is guaranteed to s members of a group is given, as also the complexity of group privacy is given. The differential privacy guarantee under k-fold adaptive composition is given (for $\varepsilon > 0$). And finally the number of papers in literature about the definition of differential privacy is given.

- Pure differential privacy is denoted as ε-differential privacy and is the result of the original work on differential privacy by Dwork in 2006 [16].
- Approximate differential privacy, denoted as (ε, δ)-differential privacy, guarantees that with probability at most $1 - \delta$ the privacy loss does not exceed ε [11].
- Mean Concentrated Differential Privacy (mCDP). An algorithm is (μ, τ)-mCDP if the privacy loss random variable has mean μ and if, after subtracting off this mean, the resulting (centered) random variable, ξ, is subgaussian with standard τ[1] [18]. It provides better accuracy than pure and approximate differential privacy and an improvement of $\sqrt{2}$ on utility/privacy trade-off in any application. Please note that mCDP is not closed under post-processing.
- Zero Concentrated Differential Privacy (zCDP) formulates concentrated differential privacy in terms of α-Rényi divergence between the distributions obtained by an algorithm running on neighboring inputs and promises sharper quantitative results than mCDP and provides the establishment of lower bounds [11]. The definition states that the α-Rényi divergence between the distributions of two neighboring databases does not exceed $\xi + \rho\alpha$, with $\alpha \in e \approx 2.718$.

2.2.2 Composition

Composition is one of the qualitative properties of differential privacy. It can be used for modular design of complex private mechanisms using simpler ones, repeated use of the same mechanism on the same database and model the interaction between many different privacy

[1] A random variable X is subgaussian with standard τ for a constant $r > 0$ if $\forall\lambda \in \mathbb{R} : E[e^{\lambda \cdot X}] \leq e^{\frac{\lambda^2 \cdot \tau^2}{2}}$.

[2] As of 22 March 2017, retrieved via Google Scholar. "ε-differential privacy" is taken as synonym for pure differential privacy.

Table 2.1 Comparison of pure differential privacy and its relaxations

Technique	Pure	Approximate	mCDP	zCDP
Definition	(ε)	(ε, δ)	(μ, τ)	(ξ, ρ)
In terms of ε-DP	(ε)	$(\varepsilon, 0)$	$(\varepsilon \cdot (e^{\varepsilon} - 1)/2, \varepsilon))$	$(\varepsilon, 0)$ or $(0, \frac{1}{2}\varepsilon^2)$
Group privacy	$(s\varepsilon)$	$(s\varepsilon, se^{s-1}\delta)$	$(s^2 \cdot \mu, s \cdot \tau)$	$(\varepsilon \cdot s \cdot (1 + \log s), \rho \cdot s^2)$
Group privacy efficiency	$O(s)$	$O(s)$	$O(s^2)$	$O(s \log s)$
Sequential composition	$(k\varepsilon)$	$(k\varepsilon, k\delta)$	$(k\mu, k\tau)$	$(k\varepsilon, k\rho)$
Papers in literature[2]	1050	139	3	4

mechanisms [19]. The level of privacy for composition using heterogeneous mechanisms depends on the level of privacy of individual mechanisms and the size of composition [44].

An example is repeatedly compute the same statistic using the Laplace mechanism [17]. The average will eventually converge to the true value of the statistic and the privacy guarantee will degrade. Another example is that if one's data is used in many differentially private data releases over his/her lifetime, involving different databases and different mechanisms, it should be clear the privacy loss accumulates [19].

2.3 Intended Intermittent Transmission

Intended intermittent transmission in this thesis means that no data is send to the data holder, thus the sender does not replace the measurement with another value. However, the data holder can decide to replace missing data values to make more accurate inferences about the population.

2.3.1 Sample Mechanism

The Sample Mechanism provides user-specific privacy budgets by non-uniform sampling each measurement. If sampling occurs at the user side, this can be seen as intended intermittent transmission. It uses a ε-differential private mechanism after sampling. These two sources of randomness provide personalized differential privacy.

2.3.2 Dealing with Missing Data

One technique to deal with missing data is imputation [27], this is the process of replacing missing data with substituted values. If only one missing data point is substituted at a time, it

is called single imputation. Multiple techniques are well-known for single imputation, for example mean imputation or last observation carried forward.

Another technique is interpolation, this method creates new data points within the range of a discrete set of known data points. Because previous data points are differential private, limited inferences could be made and thus this technique is less valuable for dealing with missing data for state estimation [15].

Dynamic filtering with intermittent observations have been researched [54]. The arrival is modeled as a random process whose parameters are related to the characteristics of the communication channel.

2.3.3 Performance Measurements

The Mean Absolute Percentage Error (MAPE) is used to measure the performance of the state estimation and imputation methods. MAPE is shown in equation (2.1), with x_i the actual measurement, $\hat{x}_i$ the estimated measurement and n the number of measurements [3]. The range of MAPE is between 0 and ∞.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \tag{2.1}$$

Chapter 3

Research Methodology

Methodology is "the branch of philosophy concerned with the science of method and procedure" [1]. Research Methodology is thus the science of methods and procedures used to perform research.

3.1 Design Science

Design Science is a paradigm in research [59]. It iterates over two activities: designing and investigation of an artifact in a context. The first activity is associated with a design problem that call for a change in the real world that require an analysis of the actual or hypothetical stakeholder goals. The latter ask for knowledge about the world it is and is described as a single truth. The activities take place in the engineering cycle and empirical cycle respectively.

Artifacts that are the subject of study are designed to interact with a problem context in order to improve something in that context. An artifact is something created by people for some practical purpose. Examples are algorithms, methods, techniques and conceptual frameworks. If the researchers works in a context with a higher-level design or engineering cycle, the research is utility-driven. Otherwise, it is an exploratory research project.

Design science of Wieringa supports making generalizations through analogic inferences.

[1] http://www.thefreedictionary.com/methodology. Retrieved 20 February 2017.

3.2 Research Goal

To frame a research, it is needed to define a research goal [59]. The research goal should be distinguished from the stakeholder goals. The research goal of this thesis is a knowledge goal, as it wants to ask the world for knowledge. It is stated as:

> Learn about the effect of adding intended intermittent transmission of differentially private smart meter measurements on state estimation accuracy and level of privacy in distribution power grids.

This is a form of validation. The goal of validation is "to predict how an artifact will interact with its context, without actually observing an implemented artifact in a real-world context" [59]. It is is experimental as artifact prototype is exposed to various scenarios presented by a model of the context to see how it responds.

3.3 Justification for Design Science

This thesis would like to answer a knowledge goal, without being in the context of a higher-level design or engineering cycle. As no knowledge is (yet) available on the combination of intermittent transmission, the empirical cycle can be used to answer the knowledge goal. Because the research aims for generalization, design science of Wieringa is chosen over just following the empirical cycle.

3.4 Empirical Cycle

The empirical cycle is a rational way to answer scientific knowledge questions [59]. Wierenga presents a checklist to help find justifiable answers to these questions.

1. **Research problem analysis**: Framing the research problem. What is exactly the problem that needs to be solved?
2. **Research design and inference design**: Designing the research setup. How to draw conclusions from data generated?
3. **Validation of research and inference design**: Validating whether the research setup and inference design match. Three kinds of validity questions about a research design can be asked:

- Inference support: To what extent does the research setup support the planned inferences?
- Repeatability: Is the design specified in such a way that competent peers can repeat the research?
- Ethics: Does the treatment of people respect ethical norms?

4. **Research execution**: Execution of the research itself.
5. **Data analysis**: The data generated by the research and is analyzed using the inference design.

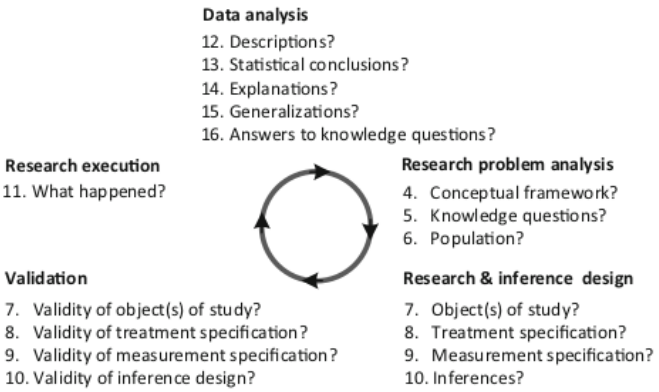


Fig. 3.1 The empirical cycle [59].

3.5 Stakeholders

A stakeholder of a problem is "a person, group of persons or institutions affected by treating the problem" [59]. In exploratory research, it is useful to think about potential stakeholders that have an interest in the research result. The functional beneficiary and their desires are:

- **Network operator** desires to enhance the observability of distribution grids.
- **Residential consumer** has the desire to keep its power consumption private. By providing intermittent power demands, instead of all measurements, the consumer is able to personalize its privacy budget. It may receive discount on its electricity bill by participating in enhancing the observability of the distribution grid.

The desire of having the capability to perform fully accurate state estimation of distribution grids is in economical conflict with the desire to not disclose any information about power consumption. This is due to the high costs of Phase Measurement Units (PMU) that need to be placed at MV/LV substations for better observability.

Other potential stakeholders are:

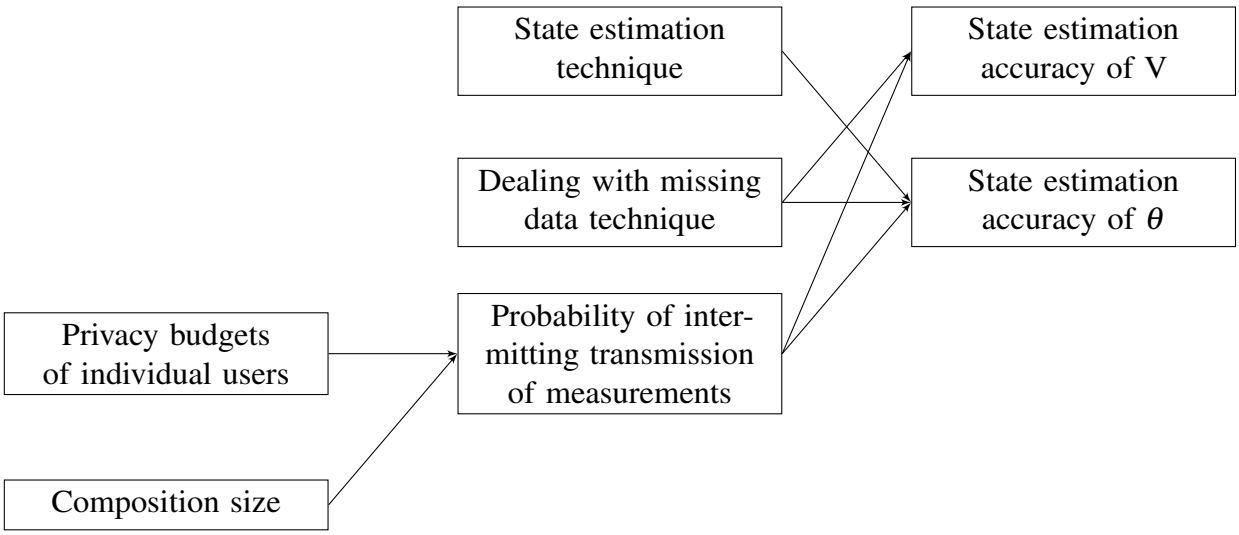
- **Manufacturer of smart meters** is a supplier and developer. They have to deliver the components of the artifact and build the system.
- **Hacker** is a threat agent and wants to compromise the integrity of the system. One desire is trying the make inferences about households such as household size, hours at work and appliances installed.

3.6 Conceptual Framework

Conceptual framework is a set of definitions of concepts that are often called constructs. They can be used to frame design and knowledge problems. Constructs are used to define the structure of the artifact and its context.

The conceptual framework is based on the chapter 2 (Preliminary Literature Research) and can be seen in figure 3.2. It uses architectural structures and support the case-based research method uses in this thesis.

Fig. 3.2 Conceptual Framework.



Construct validity is defined as the degree to which the application of constructs to phenomena is warranted with respect to the research goals and questions. Validity requirements

are adequate definition, classification of instances unambiguously, capturing the intendedly meaning of indicators and avoiding bias in the measurement methods. To be able to measure the constructs, the constructs need to be operationalization. Operationalization of a property is "a measurement procedure by which evidence for the presence of the property can be established". The constructs in the conceptual framework are operationalized below.

- **State estimation accuracy** of V and θ (variable): the accuracy of state estimation relative to the true state. It is measured using MAPE, as it is a popular measure for state estimation and can be used to compare the performance of this treatment to other treatments that provide differential privacy.
- **Level of privacy under composition** (variable): the level of privacy that is guaranteed under composition and is measured using (ε, δ)-differential privacy.
- **Dealing with missing data technique** (entity): the technique that deals with missing data.
- **Probability of intermitting transmission of measurements** (variable): the probability that each user will not send their measurement to guarantee their privacy budget.
- **Composition size** (variable): the number of measurements under which differential privacy is guaranteed to the user.
- **Privacy budgets of individual users** (variable): the level of privacy that is guaranteed for each individual user under the composition size and is measured using ε-differential privacy.

Pure differential privacy, which is denoted as ε-differential privacy, is chosen to quantify privacy over relaxations of differential privacy. This is due to the fact that pure differential privacy is used by the sample mechanism to provide individual privacy budgets to users and its popularity in literature.

3.7 Research Questions

The knowledge goal can be refined into knowledge questions. Each knowledge question can be classified as an effect, trade-off, sensitivity or requirements satisfaction question. Requirements satisfaction questions are neglected, as it is an exploratory research project and requirements analysis as part of the design cycle is not performed.

Effect questions: ask what effect an artifact in a context has.

1. What is the effect of the sample mechanism on state estimation accuracy?

Trade-off questions: ask what are the difference between effects of different artifacts in the same context.

2. What is the effect of different state estimation and imputation techniques on state estimation accuracy?
3. What is the effect of varying personalized privacy budgets on state estimation accuracy?
4. What is the effect of guaranteeing the privacy budgets under composition on state estimation accuracy?

Sensitivity questions: ask what happens if the context is changed.

5. What is the effect of the number of participating smart meters in low-voltage areas on state estimation accuracy?
6. What is the effect of changing the day of the week and month of the year on state estimation accuracy?

Chapter 4

Literature Research

4.1 Differential Privacy

Differential privacy provides "a systematic approach for responding to statistical queries on stochastic databases while preserving the privacy of individuals" [24]. It is a relatively new notion of privacy [20]. The definition guarantees that the presence or absence of an individual will not influence the output of the algorithm [39], thus all outputs are insensitive to any individual's data. There is a trade-off between utility and privacy. In many cases, accurate information can be provided while ensuring high levels of privacy [16]. The definition of differential privacy, referred to as pure differential privacy, is given as [34]:

Definition 1 (Pure Differential Privacy)**.** A randomized algorithm A is ε-*differentially private* if for any two data sets D_u and D_v differing on one element, and for all $R \subset Range(A)$,

$$Pr[A(D_u) \in R] \leq exp(\varepsilon) \cdot Pr[A(D_v) \in R]. \tag{4.1}$$

A randomized algorithm is an algorithm that uses a degree of randomness. The outcome of such an algorithm is a random variable. R is an image of the algorithm A, which is the subset of output domain of A. In the definition, the privacy budget ε is publicly known and controls the level of privacy [30]. A high ε provides weaker privacy. When ε is small, $e^{\varepsilon} \approx 1 + \varepsilon$

4.1.1 Composition and Post-processing

Differential privacy is preserved under post-processing [21, 16]. A data analyst can not compute a function of the output of a differential private mechanism and make it less differentially private. This implies that the privacy loss can not be increased under the

formal or by simply sitting in a corner and thinking about the output of the mechanism. This is formally stated in theorem 1, which states that a data-independent mapping g with a ε-differentially private mechanism $\mathscr{M}$ is also ε-differentially private.

Theorem 1 (Post-processing)**.** Let $\mathscr{M} : \mathbb{N}^{|X|} \to R$ be a ε-differentially private query and g : $R \to R'$ be an arbitrary deterministic mapping. Then,

$$g \circ \mathscr{M} : \mathbb{N}^{|X|} \to R' \tag{4.2}$$

is also ε-differentially private.

Composition is a qualitative property of differential privacy and ensures that the cumulative privacy loss can be quantified over k computations [16]. Sequential composition is used to quantify the privacy loss over multiple queries where the output is assumed to be correlated and is very pessimistic about the privacy loss. Parallel composition, on the other hand, is performed on disjoint subsets of the dataset [38]. An example where parallel composition can be used is a disjoint subset of the data by break the dataset into left/right-handed and hair color [1]. An example of sequential composition is the composition of multiple computations over time to calculate the sum of power loads [21].

The differential privacy guarantee under sequential composition using homogeneous mechanisms is formulated in theorem 2 [19], whereas the parallel composition is described in theorem 3. Homogeneous mechanisms share the same privacy budget ε, whereas heterogeneous mechanisms allow for different privacy budgets for each mechanism.

Theorem 2 (Sequential Composition Theorem for Homogeneous Mechanisms)**.** The family of ε-differentially private mechanisms satisfies $k\varepsilon$-differential privacy under k-fold sequential composition.

Theorem 3 (Parallel Composition Theorem for Homogeneous Mechanisms)**.** The family of ε-differentially private mechanisms satisfies ε-differential privacy under parallel composition.

Generally, the pure differential privacy definition is given for a pair of databases that differ one row and thus guarantees differential privacy to one participant that submitted his or her information. Group privacy bounds the privacy loss for pairs of databases that differ in the data in the case of s individuals [18], such as members of a family.

Theorem 4 (Group Privacy)**.** If mechanism A is an ε-differential private, then for all pairs of databases $x, x' \in \mathbb{X}^n$, $A(x)$ and $A(x')$ are $s\varepsilon$-indistinguishable for $s = d(x,x')$. $d(x,x')$ is the Hamming distance between x and x'.

[1] http://dimacs.rutgers.edu/~graham/pubs/slides/privdb-tutorial.pdf. Retrieved 5 June 2017

4.1.2 Personalized Differential Privacy (PDP)

The definition of pure differential privacy (given in definition 1) guarantees an uniform level of privacy for all users [30]. Data privacy is however a "personal and multifaceted concept" and users have different privacy preferences. Each user's privacy requirement is captured in the privacy specification as defined in definition 2. The privacy requirement can be interpreted as the privacy budget parameter ε in pure differential privacy.

Definition 2 (Privacy Specification)**.** A privacy specification is a mapping $\Phi : \mathscr{U} \to \mathbb{R}_+$ from users privacy preferences, where a smaller value represents a stronger privacy preference. The notion Φ^u is used to denote the privacy preference corresponding to user $u \in \mathscr{U}$.

The notion personalized differential privacy (PDP) is used to guarantee each user's privacy budget and is given in definition 3. A privacy specification is described as a set of ordered pairs for convenience in [30], such that $\Phi := \{(u_1, \varepsilon_1), (u_2, \varepsilon_2), \ldots\}$.

Definition 3 (Personalized Differential Privacy (PDP))**.** In the context of a privacy specification Φ and a universe of users $\mathscr{U}$, a randomized mechanism $\mathscr{M} : D \to R$ satisfies Φ-personalized differential privacy (or Φ-PDP), if for every pair of neighboring datasets $D, D' \subset \mathscr{D}$, with $D \sim^t D'$, and for all sets $O \subseteq R$ of possible outputs,

$$Pr[\mathscr{M}(D) \in O] \leq e^{\Phi^u} Pr[\mathscr{M}(D') \in O], \tag{4.3}$$

where $u \in \mathscr{U}$ is the user associated with type t, and Φ^u denotes u's privacy preference.

In a setting where ε-differential privacy is guaranteed to users, it is also personalized differentially private where each user's privacy requirement is set to ε:

Theorem 5 (Differential Privacy Implies PDP)**.** Let $\mathscr{U}$ denote a universe of users and let D denote the associated universe of tuples. Any mechanism $\mathscr{M} : D \to R$ that satisfies ε-differential privacy also satisfies Φ-PDP, with privacy specification $\Phi = \{(u, \varepsilon) | u \in \mathscr{U}\}$

Also, the qualitative property of composition holds for personalized differential privacy. The privacy that can be afforded to a user degrades when multiple queries are run on the same data.

Theorem 6 (Composition Theorem for Personalized Differential Privacy)**.** Let $\mathscr{M}_1 : D_1 \to R$ and $\mathscr{M}_2 : D_2 \to R$ denote two mechanisms that satisfy PDP for Φ_1 and Φ_2, respectively. Let $\mathscr{U}_1$ and $\mathscr{U}_2$ denote the associated universes of users. Finally let $D_3 = D_1 \cup D_2$. Then, for any $D \subset D_3$, the mechanism $\mathscr{M}_3(D) = g(\mathscr{M}_1(D \cap D_1), \mathscr{M}_2(D \cap D_2))$ satisfies Φ_3-PDP, where $\Phi_3 = (\{(u, \Phi_1^u + \Phi_2^u) | u \in \mathscr{U}_1 \cap \mathscr{U}_2\}) \cup \{(v, \Phi_1^v | v \in \mathscr{U}_1 \setminus \mathscr{U}_2)\} \cup \{(w, \Phi_2^w | w \in \mathscr{U}_2 \setminus \mathscr{U}_1)\}$, and g is an arbitrary function of outputs of $\mathscr{M}_1$ and $\mathscr{M}_2$.

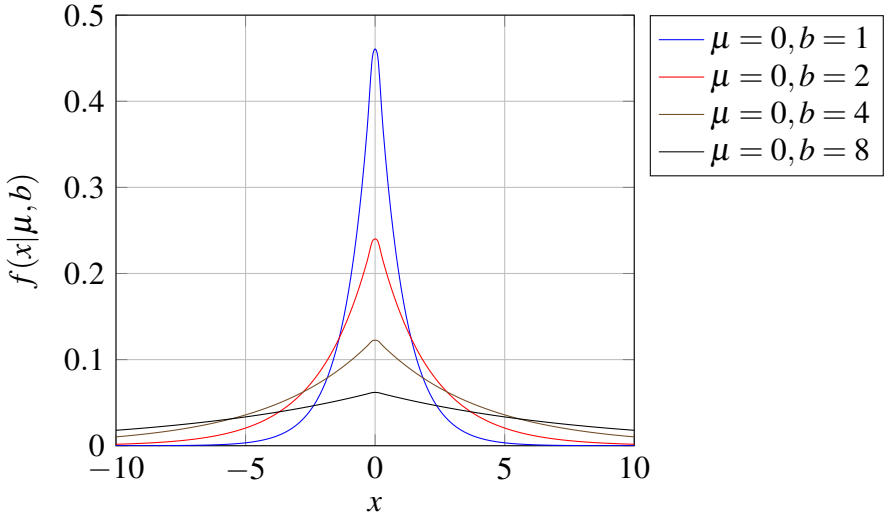


Fig. 4.1 Probability Density Function of Laplace Distribution

4.2 Differential Private Mechanisms

4.2.1 Laplace Mechanism

The Laplace mechanism is a popular differential private mechanism that adds Laplace noise to the query result with mean $\mu = 0$ and variance $2\lambda^2$ [16]. The parameter λ depends on the sensitivity $S(f)$ of the query f and the privacy budget ε. The Laplace distribution for several variances can be seen in figure 4.1. The mechanism is approximately optimal in high privacy regimes (where $\varepsilon \to 0$) using pure differential privacy [25]. It is defined as [1]:

Theorem 7 (Laplacian Mechanism)**.** For all $f : D \to R$, the following mechanism A is ε-differential private: $A(D) = f(D) + L(S(f)/\varepsilon)$, where $L(S(f)/\varepsilon)$ is an independently generated random variable following the Laplace distribution and $S(f)$ denotes the global sensitivity of f.

Where sensitivity is defined as:

Definition 4 (Sensitivity)**.** For any function $f : D \to \mathbb{R}^d$, the sensitivity of function f is:

$$S(f) = \Delta f = \max_{D,D'} ||f(D) - f(D')||_1 \tag{4.4}$$

The infinite divisible property of a distribution states that a distribution X is equal to the sum of n independent identically distributed random variables [1], it is defined in theorem 8 (using the Gamma distribution) and is used to significantly reduce noise added to the sum of measurements at a bus while preserving the differential privacy guarantee to users [1]:

Theorem 8 (Infinite Divisibility of Laplace Distribution)**.** Let $L(\lambda)$ denote a random variable which has a Laplace distribution with PDF $f(x,\lambda) = \frac{1}{2\lambda}e^{\frac{|x|}{\lambda}}$. Then the distribution of $L(\lambda)$ is infinitely divisible. Furthermore, for every integer $n \geq 1, L(\lambda) = Pni = 1[G1(n,\lambda) - G2(n,\lambda)]$, where $G1(n,\lambda)$ and $G2(n,\lambda)$ are i.i.d. random variables having gamma distribution with PDF $g(x,n,\lambda) = \frac{(1/\lambda)^{1/n}}{\Gamma(1/n)}x^{\frac{1}{n}-1}e^{-x/\lambda}$ where $x \geq 0$.

4.2.2 Sample Mechanism

The Sample Mechanism is able to personalize the privacy budget of users [30]. It can be used to achieve better utility than the Laplace Mechanism. It uses two independent sources of randomness: non-uniform random sampling at the tuple level and uniform randomness by invoking a ε-differential privacy mechanism. The probability of inclusion of a tuple π_x in the sampling part depends on the personal privacy preference Φ^u and the global threshold t. Sampling introduces uncertainty about which tuples are included in the result. The Sample Mechanism works with every differentially private mechanism DP^f.

The two sources of randomness add two types of errors to the statistic function f (in this thesis the sum). The threshold provides some balance between these types of errors, as a small threshold results in more tuples be sampled but higher error due to the differentially private mechanism DP_t^f. By setting the threshold $t = max_u\Phi^u$, each user receives precisely the required amount of privacy.

Definition 5 (Sample Mechanism)**.** Consider a function $f : \mathscr{D} \to R$, a dataset $D \subset \mathscr{D}$, and a privacy specification Φ. Let $RS(D,\Phi,t)$ denote the procedure that independently samples each type $x \in D$ with probability

$$\pi_x = \left\{ \begin{array}{ll} \frac{e^{\phi^{x_{\mathscr{U}}}}-1}{e^t-1}, & \text{if } \phi^{x_{\mathscr{U}}} < t \\ 1, & \text{otherwise} \end{array} \right\} \tag{4.5}$$

where $\Phi^{x_{\mathscr{U}}}$ denotes the privacy preference of the user associated with type x, and $min_u\phi^u \leq t \leq max_u\Phi^u$ is a configurable threshold. The Sample Mechanism is defined as:

$$S_f(D,\Phi,t) = DP_t^f(RS(D,\Phi,t)) \tag{4.6}$$

where DP_t^f is any t-differentially private mechanism that computes the function f.

4.3 Additive Homomorphic Encryption

An homomorphic encryption scheme allows arithmetic operations to be performed on ciphertexts [36]. With an additive homomorphic encryption scheme, summation of encrypted values without decrypting individual values can be performed. Several schemes that provide additive homomorphic encryption are proposed, such as Paillier's scheme and CaTsMy scheme [1, 36]. In this thesis, the CaTsMy scheme is used and described. Another possibility is the approach described in [35]. Discussion of which additive homomorphic encryption scheme is used has been neglected, as it is not modeled in this thesis and therefore only showing the opportunity is sufficient.

4.3.1 CaTsMy Scheme

The CaTsMy scheme is a simple and practical additive homomorphic encryption scheme [1]. The name refers to the authors of the paper. The algorithm is described below.

CaTsMy Scheme Algorithm

Encryption: The message that will be send is called the plaintext. The keystream is used in combination with the plaintext to produce the (encrypted) ciphertext that is send to the aggregator. The shared keystream is generated by two nodes as described in the protocol section.

1. Represent the (plaintext) message m as integer $m \in [0, M-1]$ where M is a large integer.
2. Let k be a randomly generated keystream, where $k \in [0, M-1]$.
3. Compute the ciphertext $c = Enc(m,k,M) = m + k(modM)$

Decryption: To obtain the plaintext after encryption, the keystream needs to be subtracted from the ciphertext.

1. $Dec(c,k,M) = c - k(modM)$

Addition of ciphertexts: The property of additive homomorphic encryption is used to create an encrypted sum of measurements that can only be decrypted using all measurements and pairwise keys.

1. Let $c_1 = Enc(m_1,k_1,M)$ and $c_2 = Enc(m_2,k_2,M)$
2. For $k = k_1 + k_2$, $Dec(c_1 + c_2,k,M) = m_1 + m_2$

4.3.2 Protocol

A protocol proposed in [1] is able to send differential private measurements securely using the CaTsMy scheme and an infinite divisible distribution.

System Setup

It requires the establishment of pairwise keys between each pair of nodes inside a cluster (with cluster size N), this can be achieved using Diffie-Hellman key exchange. Each node generates a private and public key. The public key is send to the other nodes through the aggregator. Using a public key of a node, other nodes can securely communicate with this node by encrypting the message with the public key. The node that receives the message encrypted with this public key can decrypt the message using his private key.

1. Each v_i sends a self-signed DH component $g^{c_i}(mod p)$ and its certificate $Cert_i$ to the aggregator. The parameter c_i is a secret and g and p are public.
2. Every time a new node is deployed, the supplier broadcasts the list of (id_i, g^{c_i}, $Cert_i$), with id_i the identity of node v_i.
3. Node v_i can compute a pairwise key $K_{i,j}$ with any other node v_j by computing $g^{c_i \cdot c_j}(mod p)$.

Node

1. Each node v_i calculates
$$\hat{X}_t^i = \mathscr{M}(X_t^i) \tag{4.7}$$
where $\mathscr{M}(x)$ is a differential private mechanism.
2. Each data with added noise $\hat{X}_t^i$ is encrypted using the CaTsMy homomorphic encryption scheme, denoted by $Enc(\hat{X}_t^i)$.
3. Each node v_i selects l other nodes at random, such that v_i selects v_j and the other way around. They generate a common dummy key k from their pairwise key $K_{i,j}$. Then, node v_i and v_j add k and $-k$ respectively to $Enc(\hat{X}_t^i)$.
4. $Enc(\hat{X}_t^i)$ is sent to the aggregator.

Aggregator

1. The aggregator receives $Enc(\hat{X}_t = \sum_{i=1}^{N} Enc(X_t^i)$.

2. The decrypted value is obtained by subtracting the keystream of the nodes:

$$Dec(\hat{X}_t) = \sum_{i=1}^{N} Enc(\hat{X}_t^i) - \sum_{i=1}^{N} K_i' = \sum_{i=1}^{N} \hat{X}_t^i \quad (mod\Delta) \tag{4.8}$$

4.3.3 Robustness

It is assumed that N nodes participate in the protocol. However, if some nodes can not participate, differential privacy can not be guaranteed as the sum of the noise will not add up to and the aggregator can not decrypt the value as the dummy-keys will not cancel out. An advanced approach is developed to tackle non-responding nodes [1], this approach requires two-way communication.

1. Each node adds a secret random value to its encrypted value in the first round before sending it to the aggregator:

   $Enc(\hat{X}_t^i) = \hat{X}_t^i + K_i' + \sum_{j=1}^{l} dkey_{i,ind_i[j]} + C_i(mod\Delta)$

2. Then, the aggregator asks all nodes for their random keys and missing dummy keys by sending their id's.

3. Each node verifies whether the nodes listed by the aggregator is in the participation node list.

4. The aggregator finally subtracts all received random values from $Enc(\hat{X}_t^i)$ which results in the desired sum $\sum_{i=1}^{N}(\hat{X}_i)$.

4.4 State Estimation

State Estimation can be performed using static or dynamic filtering. Static filtering does not use any preceding information [58], in contrast to dynamic filtering where consecutive and uncorrelated sets of measurements varying in time are used. Static filtering is known for its simplicity and fast convergence and is used in control centers around the world.

4.4.1 Weighted Least Squares

The Weighted Least Squares (WLS) state estimator is a classical state estimator widely used by power system control centers [55, 58]. It is an iterative linearizion based on the Newton-Raphson method. The objective of WLS is given in equation (4.9).

$$\begin{aligned} \min \quad & (z - f(x))^T W (z - f(x)) \\ \text{s.t.} \quad & z = f(x) + e \end{aligned} \tag{4.9}$$

Where z is the measurement vector, f is the vector of nonlinear measurement functions, x is the state and e is the measurement error vector and W is the weight measurement matrix.

Even if the initial guess is not good, the estimate will generally converge. The gradient of the estimated state at each iteration is given by:

$$\Delta \hat{x} = (F^T W F)^{-1} F^T W \Delta z \tag{4.10}$$

Where F is the $m \times n$ Jacobian matrix of f with elements $F_{ij} = \frac{\partial f_1}{\partial x_j}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

4.4.2 Iterative Recursive Weighted Least Squares

The Iterative Recursive Weighted Least Squares method is a variant of WLS [55]. Instead of assigning the same weight to each measurement for all iterations, the IRWLS method changes the weights of individual measurements at each iteration. Measurements with a large residual will get a reduced weight. The objective of the IRWLS method is:

$$\begin{aligned} \min \quad & \textstyle\sum_{i=1}^{m} (z_i - f_i(x))^2 W_i \\ \text{s.t.} \quad & z = f(x) + e \end{aligned} \tag{4.11}$$

Where z is the measurement vector, f is the vector of nonlinear measurement functions, x is the state and e is the measurement error vector and W_i is the weight measurement vector.

Equation (4.11) has been rewritten in the same formulation as the WLS state estimator:

$$\begin{aligned} \min \quad & (z - f(x))^T W (z - f(x)) \\ \text{s.t.} \quad & z = f(x) + e \end{aligned} \tag{4.12}$$

Where W is the weight matrix $diag(w_1, \ldots, w_m)$. The elements in the initial weight matrix is set to $w_i = 1/\sigma_i^2$.

Measurements are normalized relative to their standard deviations:

$$s = \frac{z}{\sigma} = [\frac{z_1}{\sigma_1}, \cdots, \frac{z_m}{\sigma_m}] \tag{4.13}$$

The normalized residual is calculated, with $h(x) = f(x)/\sigma$ as:

$$R_i = |s_i - h_i(x)| \tag{4.14}$$

The diagonal elements of weight matrix W_i are modified, with α the iteration count, as:

$$W_i^{\alpha} \propto \frac{1}{R_i^{\alpha}} \tag{4.15}$$

In order to avoid convergence problems due to very small residual errors, the range of weights is limited to 0.001 and 1 [2]. This is formulated in equation (4.16) using $R_{max}^k = \max R_i^k$.

$$W_i^k = \left\{ \begin{array}{ll} 0.001 R_{max}^k / R_i^k, & \text{if } R_i^k > 0.001 R_{max}^k \\ 1, & \text{else} \end{array} \right\} \tag{4.16}$$

4.4.3 Extended Kalman Filter

The Kalman Filter propagates the current estimated mean and covariance through the system dynamics to generate a prior distribution for the next state estimate [12]. This prior is used in conjunction with the likelihood function from the measurements to form a posterior used in estimation. The accuracy is reduced by the linear approximation of EKF.

The Extended Kalman Filter (EKF) calculates a Jacobian matrix to linearize the measurement equation (4.17) [58].

$$y_k = h(x_k) + r_k \tag{4.17}$$

With y_k the measurement vector, $h(x_k)$ the non-linear equations modeling y_k as function of bus voltages, angles and network parameters and r_k the Gaussian white noise of the measurements at time k.

The EKF involves two steps [26], namely prediction and correction steps. The initial state at $k = 0$ needs to be set accordingly to the system.

Step 1: Predict

The + (plus) symbol denotes an a posteriori estimate.

$$P_k^- = F_{k-1} P_{k-1}^- F_{k-1}^T + L_{k-1} Q_{k-1} L_{k-1}^T \tag{4.18}$$

$$\hat{x}_k^- = f_{k-1}(\hat{x}_{k-1}^+, u_{k-1}, 0) \tag{4.19}$$

Where F and L are the process Jacobians at step k and are shown in equations (4.20) and (4.21). P is the co-variance matrix of the state estimation error (also called gain factor matrix) and Q is the process noise covariance at step k.

$$F_{k-1} = \left.\frac{\partial f_{k-1}}{\partial x}\right|_{\hat{x}_{k-1}^+} \tag{4.20}$$

$$L_{k-1} = \left.\frac{\partial f_{k-1}}{\partial w}\right|_{\hat{x}_{k-1}^+} \tag{4.21}$$

Step 2: Correction

The measurement update of the Kalman gain, state estimation and estimation-error covariance is calculated:

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + M_k R_k M_k^T)^{-1} \tag{4.22}$$

$$\hat{x}_k^+ = \hat{x}_k^- + K_k[y_k - h_k(\hat{x}_k^-, 0)] \tag{4.23}$$

$$P_k^+ = (I - K_k H_k) P_k^- \tag{4.24}$$

Where H and M are the measurement Jacobians at step k and shown in equations (4.25) and (4.26).

$$H_k = \left.\frac{\partial h_k}{\partial x}\right|_{\hat{x}_k^-} \tag{4.25}$$

$$M_k = \left.\frac{\partial h_k}{\partial v}\right|_{\hat{x}_k^-} \tag{4.26}$$

4.4.4 Unscented Kalman Filter

The Unscented Kalman Filter (UKF) is a discrete-time recursive filter that is able to solve estimation problems in the form of equations (4.27) and (4.28) [58]. It is based on the

unscented transformation and Kalman filter theories. It copes with non-linear models by propagating a statistical distribution of the state through the non-linear equations.

$$x_k = f(x_{k-1}, k-1) + q_{k-1} \tag{4.27}$$

$$y_k = h(x_k, k) + r_k \tag{4.28}$$

With x the state vector, y the measurement vector, q_{k-1} and r_k are the system and measurement Gaussian noises with zero mean and uncorrelated covariance matrices Q and R respectively [58]. f and h are non-linear functions that represents the system and measurement models in terms of state variables and other system inputs.

The UKF is easier to implement than the EKF, as it does not require computing any derivative or Jacobian [58]. It provides approximations that are accurate to the third order for the Gaussian distribution for any non-linear system and to the second order for any non-Gaussian distributions. Generally, the UKF is more accurate than EKF by not linearizing the model. A drawback is that it requires much more computing power.

The UKF consists of three steps at each iteration k: sigma points calculation, Kalman filter state prediction and finally state correction.

Step 1: Sigma Points Calculation

First, the set of $2n+1$ sigma points is needed to be created.

$$X_{k-1} = [x_{k-1} \cdots x_{k-1}] + \sqrt{c}\begin{bmatrix} 0 & \sqrt{P_{k-1}} & -\sqrt{P_{k-1}} \end{bmatrix} \tag{4.29}$$

Where $c = n + \lambda$, x the state vector at time $k-1$ and covariance matrix P_{k-1}. The initial state vector and covariance matrix have to be defined at $k = 0$ according to prior knowledge of the system.

Step 2: Predict

The calculated sigma points in step 1 are evaluated by the state-update function:

$$\hat{X}_k^i = f(X_{k-1}^i, k-1) \text{ for } i = 0, \ldots, 2n \tag{4.30}$$

Where X_{k-1}^i is the (i+1)th column of matrix X_{k-1} and the resulting X_k is the $n \times (2n+1)$ matrix containing the propagated sigma points. Next, the predicted state vector x_k^- and covariance matrix P_k^- is calculated:

$$x_k^- = \sum_{i=0}^{2n} W_i^m \hat{X}_k^i \tag{4.31}$$

$$P_k^- = \sum_{i=0}^{2n} W_i^c [(\hat{X}_k^i - x_k^-)(\hat{X}_k^i - x_k^-)] + Q_{k-1} \tag{4.32}$$

Step 3: Update

The sigma points corresponding to the predicted state mean vector and covariance matrix are calculated:

$$X_k^- = [x_k^- \cdots x_k^-] + \sqrt{c}\begin{bmatrix} 0 & \sqrt{P_{k-1}} & -\sqrt{P_{k-1}} \end{bmatrix} \tag{4.33}$$

Then, the sigma points are propagated through the measurement update function h and the mean of the propagated points μ_k is calculated:

$$Y_k^- = h(X_k^-, k) \tag{4.34}$$

$$\mu_k = \sum_{i=0}^{2n} W_i^m Y_k^{-^i} \tag{4.35}$$

Next, the measurement covariance matrix and the cross-covariance of the state and measurement are calculated:

$$S_k = \sum_{i=0}^{2n} W_i^c [(Y_k^{-^i} - \mu_k)(Y_k^{-^i} - \mu_k)^T] + R_k \tag{4.36}$$

$$C_k = \sum_{i=0}^{2n} W_i^c [(X_k^{-^i} - x_k^-)(Y_k^{-^i} - \mu_k)^T] \tag{4.37}$$

At last, the filter gain K_k, state mean x_k and the covariance P_k are calculated:

$$K_k = C_k S_k^{-1} \tag{4.38}$$

$$x_k = x_k^- + K_k[y_k - \mu_k] \tag{4.39}$$

$$P_k = P_k^- - K_k S_k K_k^T \tag{4.40}$$

4.4.5 Particle Filter

UKF has reasonable performance, although there are some drawbacks that can overcome by the particle filter [23]:

- True global approximation can not be done using UKF due to the small set of sigma points.
- Less effective with nearly deterministic systems.
- Practical implementation issues due to estimation of noise covariance matrices.
- Only application is to systems with unimodal distribution and Gaussian noise.

The designer of the particle filter algorithm can choose the number of particles that represent the posterior distribution of the estimated states. A particle filter can be applied to both Gaussian and non-Gaussian distributions.

The resampling step is one of the most important step in particle filters [23], such as systematic, multinomial, stratified and residual resampling. Systematic resampling is favored due to its resampling quality, simplicity and computational complexity. Therefore, this method is used in the algorithm.

Resampling only is performed if the number of efficient particles falls below a threshold N_{th}. The efficient number of particles N_{eff} can be calculated using equation (4.41) or its approximation $\hat{N}_{eff}$ in equation (4.42).

$$N_{eff} = \frac{N}{1 + N^2 Var(w^i_{(k|k)})} \tag{4.41}$$

$$\hat{N}_{eff} = \frac{1}{\sum_{i=1}^{N} (w^i_{(k|k)})^2} \tag{4.42}$$

Step 0: Initialization at k=0

Let the number of particles be N and $q(x(k+1)|x(k), y(k+1)) = p(x(k+1)|x(k))$. Then, particles $x^i(0)$ are drawn for $i = 1, 2, \ldots, N$ randomly from p(x(0)) and let the weight $w^i_{1|0} = 1/N$.

Step 1: Measurement update

For $i = 1, \ldots, N$ assign a weight to each particle according to:

$$w^i_{(k|k)} = w^i_{(k|k-1)} p(y(k)|x^i(k)) \tag{4.43}$$

Then, normalize the weights:

$$w^i_{(k|k)} = w^i_{(k|k)} / \sum_{j=1}^{N} w^j_{(k|k)} \tag{4.44}$$

Step 2: Systematic adaptive resampling

The systematic adaptive resampling procedure is as follows:

1. Calculate N_{eff} (using equation (4.41) or (4.42)) and set $N_{th} = N/2$. If $N_{eff} < N_{th}$, then continue, otherwise go to step 3.

2. Calculate cumulative $c_i = c_{i-1} + w^i_{k:ki} = 2^N$, $c1 = 0$ by finding the cumulative sum of the elements of $w_{(k|k)}$, i.e. $c_1, c_2, \ldots, c_N$.

3. Draw a starting sample from a zero mean, $1/N$ variance, normal distribution $\mathscr{N}$:

$$s_1 \sim \mathscr{N}(\cdot, 0, 1/N) \tag{4.45}$$

4. Do N iterations from $j = 1, \ldots, N$ calculating $s_j = s_1 + (j-1)/N$ and at each iteration find the first value on the cumulative $c_1, c_2, \cdots, c_N$ that is greater than s_j and register the corresponding index of c_i in set $\mathscr{S}$.

5. Select the particles that correspond to the elements in $\mathscr{S}$ and assign identical weight $1/N$ to these selected particles.

Step 3: Time update

Estimate the state mean according to:

$$\hat{x}(k) = \sum_{i=1}^{N} w^i_{(k|k)} x^i(k) \tag{4.46}$$

Generate predicted particles for the next iteration as:

$$x^i(k+1) \sim p(x(k+1)|x^i(k)) \tag{4.47}$$

And calculate the importance weights at time step $k+1$:

$$w^i_{(k+1|k)} = w^i_{k|k} \tag{4.48}$$

Last, set $k = k+1$ and go to step 1.

4.5 Imputation Methods

Data imputation estimates values for the identified bad and missing measurements [45]. Popular imputation methods are last observation carried forward (LOCF), mean imputation, multiple imputation, interpolation and extrapolation.

Last Observation Carried Forward

The missing data value is substituted by the last observed value [27]

Mean Imputation

The missing data value is substituted by the average of the observed values [27].

Multiple Imputation

Multiple imputation uses three steps to estimate the missing value [27]. First, a plausible multivariate distribution for missing observations is estimated. The missing missing values are replaced by values randomly drawn from this distribution. Second, this step is repeated, which results in a number of completed datasets. These datasets are analyzed using complete-data methods. Last, the datasets are combined into a single dataset.

Interpolation and Extrapolation

Interpolation estimates the missing value from previous and succeeding available values [45], while extrapolation estimates missing values beyond the range of the dataset. An example of linear interpolation using preceding values y_h and succeeding values y_j is [45]:

$$\hat{y}_i^{LI} = y_h + \frac{y_j - y_h}{x_j - x_h}(x - x_h), x_h < x_i < x_j. \tag{4.49}$$

Chapter 5

Modeling

In this chapter, the components of research design are discussed. These are the object of study, treatment and measurement [59].

5.1 Research Setup

A validation model (object of study) is used that represents entities of interests that are called targets, so that questions about the targets can be answered by studying the model [59]. A validation model consists of a model of the artifact that interacts with a model of the problem context. It needs to support descriptive, abductive and analogic inferences to be valid. Here, the assessment of the validity of inferences is meant in contrast to the validity of a treatment design during the engineering cycle.

The research method used is single-case mechanism experiments [59], which is a test of a mechanism in a single object of study with a known architecture. It is a useful experiment where the researcher "applies stimuli to the validation model and explains the response using the mechanisms internal to the model". Ethics as formulated by Wierenga do not play a role in this research, as no people will be studied.

Variables and Constructs

The measurement of a variable is the "assignment, according to a rule, of a value to the phenomenon denoted by the variable". Defining a measurement rule involves defining a measuring scale. A measurement scale is a set of values with manipulation rules and should have real-world meaning. The variables and constructs can be seen in the conceptual framework in figure 3.2. Scales are determined as:

- **State Estimation accuracy** of V and θ: Ratio

- **Level of privacy under composition**: Ratio
- **Dealing with missing data technique**: Nominal
- **Probability of intermitting transmission of measurements**: Ratio
- **Privacy budgets of individual users**: Ratio
- **Composition size**: Ratio

The nominal scale indicates a classification, while ratio scale measures "how much of a unit scale goes into the measured phenomenon" [59]. The latter has a zero on the scale.

Domestic Electricity Demand Dataset

The CREST model in Excel is used to create household load profiles [48]. The model generates a synthetic consumption profile with a one-minute resolution for an Irish household. The input of the model is the month, type of day (working or weekend day). A Markov Chain approach is used to determine the number of residents active in the household. A macro is added to he Excel file to automatically create 100 load profiles for each bus for a weekday in May. The load profiles that are generated for each bus are numbered from 1 to 100. The characteristics of the dataset can be seen in Appendix A. The dataset does not include power injections (e.g. by PV or batteries). This has also been neglected for this thesis, as no dataset (generator) is available to my knowledge that would fit the one-minute interval dataset that is used. The loads are scaled per bus to the maximum load of a bus according to [3]. No random noise is added that represent the power loss during transmission, as it would add randomness that probably has influence on the outcome when running 25 runs per experiment (Monte Carlo simulations).

A dataset comprises of 100 load profiles for each bus. A typical LV feeder can serve up to 40-50 households [8], however more than 50 households is also reported [41]. Therefore, the number of households per bus N is varied: 10, 25, 50, 75 and 100. These represent the load profiles 1 to N per bus. So if $N = 10$, the load profiles 1, 2, ..., 10 will be used. Multiple datasets are created: 12 datasets for each month during weekdays and 12 datasets for each month during weekends.

The CREST model uses appliances in the categories cold, consumer electronics+ict, cooking, wet, water heating, electric space heating and lighting. These appliances can be turned on and off, dependent on the number of active occupants, time that an appliance is active, cycle time, delay after active and time active over the year.

The algorithm (implemented in Excel) to create the data in a simplified form is:
Initial:

1. Create houses according to number of residents distribution
2. Populate each household with appliances.
3. Determine initial state of household

At each time step:

1. Perform transition of number of occupants in each household
2. Trigger switch-on event

Measurement Instrument

The instrument to perform measurements is by running a (computer) simulation. The simulation is written in Python 2.7 and runs on Ubuntu 17.04. Some modules that are used are Pypower, numpy and plot.ly for creating plots. Hardware used is a Intel i5-6600K and 32GB ram. A huge difference in performance has been noticed in the state estimation techniques. The IRWLS performs best in terms of computation time with 25 iterations (it) per second, EKF runs at 23 it/s and UKF runs at 0.72 it/s.

Provenance

Data storage involves "maintaining traceability between the data and the data source" [59], this is called provenance. Also the decision who can use the data is part of data storage. As state estimation examples for power systems are rarely published on the internet, I choose to publish the source code of my thesis along with the dataset so that others can extend and replicate the work. The dataset is already available on https://github.com/rutgerprins/thesis. The source code will be disclosed when the source code has been optimized. The modified Excel file of CREST is also published in the repository conform the GPL 3.0 license of the original CREST Excel file.

5.2 Differential Private Mechanisms

5.2.1 Composition

Because the active and reactive power are the most used signatures for non-intrusive load monitoring [63], the two variables are correlated and sequential composition theorem is used to determine the required privacy budget [1]. By using sequential composition, the privacy budget of each user can be guaranteed under k-fold composition if their privacy budget of

an individual measurement is set to $\frac{\varepsilon}{k}$, as $\varepsilon = k \cdot \varepsilon_i$ according to theorem 2. Therefore, the probability of sampling tuples in the Sample Mechanism is modified to guarantee differential privacy under k-fold sequential composition for a user (original can be seen in equation (4.5)).

$$\pi_x = \begin{cases} \frac{e^{\phi^{x_{\mathscr{U}}}/k}-1}{e^{(t/k)}-1}, & \text{if } \phi^{x_{\mathscr{U}}} < t \\ 1, & \text{otherwise} \end{cases} \tag{5.1}$$

5.2.2 Sampling Mechanism

Users can be divided in three groups that represent their privacy concern [30]: conservative, moderate and liberal. The fraction of users in the conservative group, with high privacy concern, is f_C. f_M represent the medium privacy concern group moderate. The fraction of users in the liberal group is $1-(f_C+f_M)$. The default values used in their experiments were $f_C = 0.54$ and $f_M = 0.37$. The parameters where determined using a survey about privacy concern.

The individualized privacy budget of users are uniform drawn from $[\varepsilon_C, \varepsilon_M]$ for the conservative group and $[\varepsilon_M, \varepsilon_L]$ for the moderate group. The liberal group has a fixed privacy budget of ε_L. The parameters and their defaults are shown in table 5.1.

Table 5.1 Parameters of user privacy concern and default values

Group	Fraction of users	Individual privacy budget
Conservative	$f_C = 0.54$	$[\varepsilon_C, \varepsilon_M] = [0.01, 0.2]$
Moderate	$f_M = 0.37$	$[\varepsilon_M, \varepsilon_L] = [0.2, 1.0]$
Liberal	$1-(f_C+f_M) = 0.09$	$\varepsilon_L = 1.0$

The fraction of users of the conservative group is varied as in the paper [30]. The same range of values of f_C used in [30] are used: $0.1, 0.2, 0.3, 0.4, 0.5, 0.54, 0.6$, where 0.54 is the default value for f_C. The privacy budget f_C for the conservative group is varied: 0.01 (default), 0.05, 0.10, 0.20, 0.30, 0.40 and 0.50. The privacy budget f_M of the moderate group is varied: 0.05, 0.10, 0.15, 0.20 (default), 0.25, 0.30, 0.35, 0.40, 0.45 and 0.50.

The sampling mechanism uses a mechanism that provides pure differential privacy. The most popular and simple mechanism that provides this, is the Laplace mechanism [16]. Therefore, the mechanism DP_t^f represents the Laplace Mechanism.

The default threshold S is set to $t = max_u \Phi^u = \varepsilon_L = 1.0$. However to see how the Sample Mechanism compares with the Laplace Mechanism alone, the threshold is also set to the minimum: $t = min_u \Phi^u$ and the Sample Mechanism simplifies to the Laplace Mechanism. As

ε_C (the lowest privacy budget) is varied, the threshold is also varied with: 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 1.0 (default). Most users will receive a much stronger level of privacy than they require by this simplification.

5.2.3 Laplace Mechanism

The Laplace Mechanism is taken as S in the sampling mechanism. It relies on the sensitivity $S(f)$, as seen in definition 4. The sensitivity is set to the maximum load of an individual household for each bus over the simulation timespan, thus the sensitivity differs per bus. The privacy budget ε of the Laplace Mechanism has been set to the threshold t as per definition 5.

5.3 State Estimation

The state vector x is modeled as a concatenation of V and θ, without the reference voltage angle θ_1.

$$x^T = [V_1, V_2, \cdots, V_N, \theta_2, \theta_3, \cdots, \theta_N] \tag{5.2}$$

Three state estimators are modeled: IRWLS, EKF and UKF. Note that the particle filter has been neglected, as due to the required weeks of simulation (with 10k particles) it could not been completed within the time frame of the thesis and can therefore be used in further research.

- **Iterative Recursive Weighted Least Squares (IRWLS)** (static)
- **Extended Kalman Filter (EKF)** (dynamic)
- **Unscented Kalman Filter (UKF)** (dynamic)

Dealing with Missing Data

To deal with the missing data generated by the Sample Mechanism, missing data is imputed by the mean (MEAN) and last-observation-carried-forward (LOCF) strategies. Both strategies can be used by the network operator in the untrusted setting. However, for the partly and fully trusted setting the LOCF imputation technique can not be used. The non-differential private measurement z_i should only be usable for one aggregated measurement at a single time step, as the privacy budget can not be guaranteed anymore due to composition.

The first one by calculating $\hat{z} = \frac{\sum z_i}{N_t} \cdot N$, with $\hat{z}$ the approximated aggregated value, z_i the measurement of a household, N_t the number of transmitted measurements and N the number of households in a low-voltage area. With LOCF, the last transmitted measurement

of a household is used to replace the missing measurement z_i. The multiple imputation and interpolation methods are not usable due to the fact that individual measurements are encrypted and thus not accessible in the partly trusted scenario.

The sampling will only run for $t > 0$ to be able to impute a missing data point with the last observation carried forward method. With the mean strategy, additional Gamma noise is added when the distributed Laplace mechanism is used in the partly trusted scenario to hold the infinite divisibility property of the Laplace distribution (theorem 8).

Measurement functions

The measurement functions define the measurement function $h(x)$ in the EKF and UKF and the Jacobian in the IRWLS. Power injection equations are used in this thesis, as the injected power is measurable. Power flow equations are therefore neglected.
Power injection equations:

$$P_k = V_k \sum_{j \in N_k} V_j (G_{kj} \cos\theta_{kj} + B_{kj} \sin\theta_{kj}) \tag{5.3}$$

$$Q_k = V_k \sum_{j \in N_k} V_j (G_{kj} \sin\theta_{kj} + B_{kj} \cos\theta_{kj}) \tag{5.4}$$

IRWLS

The Jacobian matrix H is calculated as in equation (5.5) [43]. It is not a square matrix, but consists of (2N-1) columns with N the number of buses. Let $H_{P,\theta}, H_{Q,\theta}, H_{Q,V}, H_{Q,V}$ denote the sub-matrices of the Jacobian H with the first symbol in the suffix indicate the measurement and the second the variable on which the partial derivative is obtained. The tolerance is set to 10^{-4} as in [50].

$$H = \begin{pmatrix} H_{P,\theta} & H_{P,V} \\ H_{Q,\theta} & H_{Q,V} \end{pmatrix} \tag{5.5}$$

Extended Kalman Filter

The Extended Kalman Filter (EKF) uses the same Jacobian matrix H of the IRWLS. The process noise Q is a tuning parameter that allows for a trade off between accuracy and time lag. A higher Q means that noise measurements get more weight in the Kalman gain and accuracy may be reduced, but the lag of state changes is reduced. It is a diagonal matrix, with its diagonal elements set to 10% of the maximum state change [47]. The maximum

state change is determined by the maximum difference between two sequential time steps of the true state calculated using PyPower. The initial estimation error covariance P_0 is a diagonal matrix with its elements set to 10^{-3}. The initial state is set to the true state of the system at $t = 0$ [46]. This true state is obtained by running the Newton Raphson power flow analysis in PyPower.

The Bienaymé formula, as stated in equation (5.6), states that the variance of the sum of uncorrelated random variables is the sum of their variances [53]. The covariance for the measurement noise R is a diagonal matrix, with its elements set to the sum of the covariances of the noise added to each measurement at the household. In the base scenario (without any addition of noise), it is set to 10^{-4}. In the case of the untrusted scenario, this is $N \cdot 2(\frac{S(f)}{\varepsilon})^2$. For the (partly) trusted scenario, this is $2(\frac{S(f)}{\varepsilon})^2$. For this thesis, the noise measurement matrix R is time-invariant.

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) \tag{5.6}$$

By using the mean imputation method as described in section 5.3, the variance of the mean is given by equation (5.7) [53]. In this equation, $\bar{X}$ is the mean of i.i.d. random variables and n the number of measurements sampled from the Sample Mechanism for a bus. This means that the variance is proportional to the number of sampled tuples. This means that the measurement noise matrix of R can be optimized if the network operator has information about the privacy preferences.

$$Var(\bar{X}) = \frac{\sigma^2}{n} \tag{5.7}$$

Unscented Kalman Filter

The Unscented Kalman Filter (UKF) uses the same parameters and initial state as the EKF, although it does not use the Jacobian. Instead it uses sigma points as a statistical linearization technique [58]. Merwe Scaled sigma points are most popular and are used in the UKF [1]. The parameters which are known to be optimal for Gaussian noise are used. The spread around the mean α is set to $1e-3$. $\beta = 2$ is optimal for Gaussian noise and the secondary scale parameter κ is set to 0.

[1] http://filterpy.readthedocs.io/en/latest/_modules/filterpy/kalman/sigma_points.html

5.4 Network

The IEEE 33-bus case is used in this thesis to asses the trade-off between state estimation accuracy and the guaranteed level of differential privacy under composition [5]. The network can be seen in figure 5.1. Parameters of this network are described in Appendix B. Bus 1 is the HV/MV transformer and higher bus numbers represents MV/LV transformers.

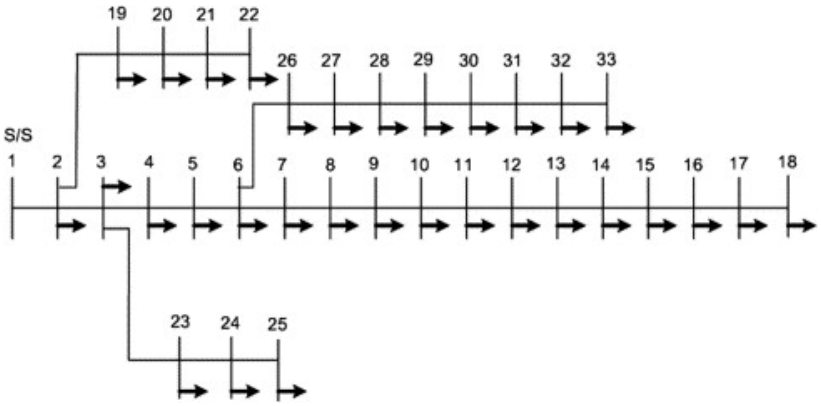


Fig. 5.1 IEEE 33-bus medium voltage network.

5.5 Scenarios

The divisibility property of the Laplace mechanism and homomorphic encryption described in the Literature Review chapter allows for multiple levels of trust in an aggregator by a household [1]. These levels of trust in the aggregator are classified into fully trusted, partly trusted and untrusted. These scenarios are compared to the base scenario which is considered as the true power flow where all measurements exist and accurate. All scenarios are summarized in table 5.2 and described in the next subsections. Encryption is required at the "partly trusted" scenario to ensure differential privacy. Please note that encryption is probably required in each setting from a security perspective.

Table 5.2 Comparison of scenarios

Trust Level	Fully Trusted	Partly Trusted	Untrusted
Individual Perturbation	-	$\Gamma(\cdot)$	$L(\cdot)$
Aggregator Perturbation	$L(\cdot)$	-	-
Encryption Required	✖	✔	✖

The scenarios consist of n households that each transmit their (real) measurement z_i to the aggregator (denoted with $\sum$ symbol) [3]. The aggregator sends the aggregated measurement

$z = \sum_{i=1}^{n} z_i$ to the operator. The operator also receives the voltage magnitude and true active/reactive power injections at bus 1. The operator then performs state estimation on the network.

5.5.1 Base Scenario

In the base scenario, all measurements at the MV/LV buses are available and accurate. Load flow analysis using PyPower[2] is performed to obtain the state of the network that is considered as true. The state $x_{hvmv}^{T} = [V, \theta]$ at the HV/MV substation (bus 1) is obtained by the network operator and used together with the aggregated measurements z_i from households at each MV/LV substation.

5.5.2 Fully Trusted Aggregator

In the fully trusted scenario, the aggregator can know the exact measurement of each household. The dwelling utilizes the $RS(\cdot)$ sampling procedure in the Sample Mechanism and thus measurements will not always be sent if the personalized privacy budget is lower than the threshold of the Sample Mechanism. After receiving all measurements, the aggregator uses the Laplace mechanism in order to complete the Sample Mechanism procedure. Personalized differentially privacy is then guaranteed to each dwelling. This can be seen in figure 5.2.

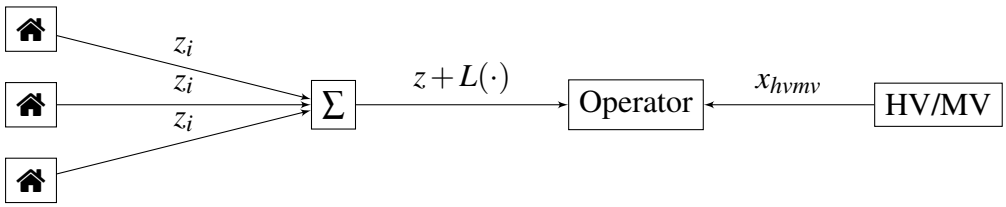


Fig. 5.2 Fully trusted aggregator scenario (household indicated by house icon and aggregator by sum icon)

5.5.3 Partly Trusted Aggregator

The partly trusted aggregator scenario utilizes the property of infinite divisibility of the Laplace distribution (see theorem 8) and homomorphic encryption in section 4.3. The dwellings utilize the $RS(\cdot)$ sampling procedure. If the measurement is sampled, the dwelling send his measurement noised to the aggregator after noising it with the Gamma distribution and encrypting it. After decrypting the sum of the noised measurements by the aggregator,

[2]https://github.com/rwl/PYPOWER

the resulting noise is not equal to the Laplace distribution due to missing measurements. Therefore, the aggregator needs to add Gamma noise for each missing measurement, so that eventually Laplace noise is added and therefore personalized differential privacy can be guaranteed.

Although the aggregator can not easily obtain the Gamma noised measurements by encryption, it (or another adversary) can decrease the level of privacy by deploying malicious nodes or lying about non-responding nodes as the Gamma distribution relies on the number of responding nodes [1]. By adding less Gamma noise than required, differential privacy can not be guaranteed to the dwellings. Therefore, this scenario is called "partly trusted".

This scenario can be seen in figure 5.3. It is modeled as the fully trusted aggregator scenario, as it is basically the same scenario but with homomorphic encryption. It is assumed the homomorphic encryption approach in section 4.3 works as described.

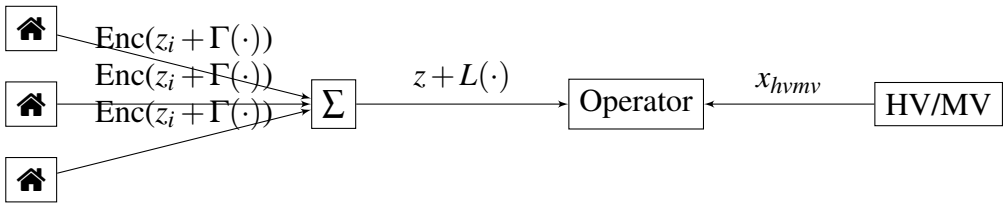


Fig. 5.3 Partly trusted aggregator scenario (household indicated by house icon and aggregator by sum icon)

5.5.4 Untrusted Aggregator

In the untrusted setting, the dwellings do not trust the aggregator as much as in the fully and partly trusted scenarios. Therefore, each dwelling utilizes the Laplace Mechanism to ensure at least t-differential privacy is guaranteed (and thus also under post-processing). Also, the sampling procedure $RS(\cdot)$ is utilized by the dwelling itself. Three possibilities for this scenario are available:

1. Each dwelling utilizes the Sample Mechanism $DP_t^f(RS(\cdot))$ and all dwellings will send their measurements using this mechanism at any time. If a measurement is not sampled, it will send $DP_t^f(0) = 0 + L(\cdot)$ to the aggregator. No trust is required by the dwelling in the aggregator.

2. Additive homomorphic encryption can be used to ensure the aggregator receives the resulting value after applying the Sample Mechanism. This means that by not sampling a measurement, the dwelling needs to send $Enc(0)$ to ensure the aggregator does not know which measurements are not sampled. This results in a sum for the aggregator

equal to $z = nL(\cdot)$, with n the number of measurements received by the aggregator. No trust in the aggregator is needed by the dwelling.

3. Intermittent transmission without encryption. The aggregator receives t-differentially private measurements from dwellings that sampled their measurements. No measurements are received if a measurement is not sampled. This means that only after summing the measurements, the Sample Mechanism $DP_t^f(RS(\cdot)) = z + nL(\cdot)$ is achieved. The notion n denotes the number of measurements received. Some trust in the aggregator for not using the individual measurements z_i is needed by the dwelling.

For this thesis, the third option is used as differential privacy is achieved partly by intermittent transmission of measurements in this thesis. With the first two options, the dwellings always send information to the aggregator. This means that the dwellings will need to have some trust in the aggregator, although their measurements are at least t-differentially private (or not available for the aggregator at all). This can be seen in figure 5.4. It results in adding the Laplace distribution n times to the aggregated measurement and therefore could result in a high error.

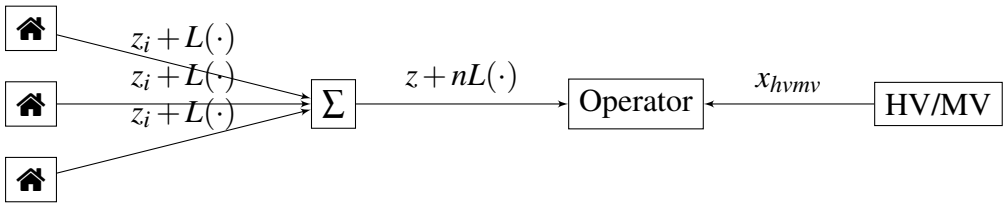


Fig. 5.4 Untrusted aggregator scenario (household indicated by house icon and aggregator by sum icon)

5.6 Simulation

Monte Carlo simulation is the generation of random objects or processes by means of a computer [32]. A purpose of Monte Carlo simulation is to sample: gather information about a random object by observing many realizations of it. The idea is to run an experiment repeatedly to obtain quantities of interest using the Law of Large numbers and other methods of statistical inference.

This thesis will use Monte Carlo simulation to obtain more accurate results by calculating the mean and standard deviation of the results. Each experiment is run for 25 times. Higher number of simulations will result in simulation times of over 48 hours and therefore not practical. Multiple elements contain randomization in this thesis:

- The allocation of privacy budgets is randomized
- The Sampling mechanism and Laplace mechanism are randomized

5.7 Inferences from Data

The validation model needs descriptive, abductive and analogic inferences to be valid [59]. The relations between each type of inference can be seen in figure 5.5.

- Descriptive inference creates a summary of phenomena by looking at the data. Wieringa describes three interpretation methods: conceptual analysis, content analysis and ground theory. In conceptual analysis, the researcher uses the conceptual framework to interpret the data. With content analysis, concepts found in the analysis are added the the conceptual framework. And with ground theory, the researcher tries not to use the conceptual framework.
- Abductive inference is inference to the best explanation(s). Three different types of explanations can be distinguished that can explain phenomena: causal, architectural and rational. A causal explanation has the form "Y changed because, earlier, a variable X changed in a particular way". An architectural explanations has the form "phenomenon E happened in the object of study because components $C_1, \ldots, C_n$ of the object of study interacted to produce E". And in rational explanations a phenomenon happened as the actor wanted to achieve a goal.
- Analogic inference is generalization by similarity of an explanation and is done after abductive inference. It is done by induction over multiple positive and negative cases and is called analytical induction.

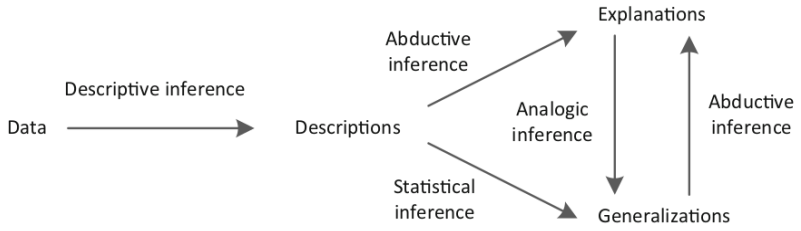


Fig. 5.5 Road map of inferences [59].

For single-case mechanism experiments, descriptive inference, architectural abductive inferences and analogic inferences are made [59]. The inferences that will be made to answer the research questions in section 3.7 are described in the following subsections.

5.7.1 Descriptive Inferences

Descriptive inferences are created from the data collected during research execution. All descriptive inferences will be graphs. Data used for these graphs can be seen in Appendix C. The graphs will be either box plots or line graphs with error bars, as Monte Carlo simulations are run. On the y-axis, the state estimation accuracy measured in mean MAPE is shown. The x-axis denotes the treatment: scenario, privacy budgets, dealing with missing data technique and/or state estimation technique. The box plots include a dotted diamond with a horizontal line in the middle. The horizontal line represents the mean and the top and bottom corners represent the standard deviation.

The first experiment is about the effects of the state estimators IRWLS, EKF and UKF on state estimation accuracy for all scenarios and imputation techniques. This means that the three state estimators are examined for five treatments: base scenario, untrusted scenario for both imputation techniques and the (partly) trusted scenario for both imputation techniques.

The other experiments run for both the (partly) trusted and untrusted scenarios are shown in table 5.3. The parameter t is the treshold of the Sample Mechanism. The privacy budgets for the conservative and moderate group are denoted by ε_C and ε_M respectively. The fraction of users in the conservative group is denoted as f_C and the composition size as k. The number of participating smart meters in an area is denoted as N. The variable m stands for month and d for day of the week, that are both inputs for the dataset generator. All graphs will be line graphs, except for the graphs with on the x-axis the month that will be box plots. The default value means that this value is used in all experiments, unless the values of the parameter are varied in th experiment.

Table 5.3 Planned experiments

Parameter	Default	Variations
t	1.0	0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 1.0
ε_C	0.01	0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50
ε_M	0.20	0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50
f_C	1	0.10, 0.20, 0.30, 0.40, 0.50, 0.54, 0.60
k	1	1, 2, 5, 15, 30, 60
N	100	10, 25, 50, 75, 100
m	5	1, 2, ..., 12
d	1	1, 5

5.7.2 Abductive Inferences

Abductive inferences can be made from the descriptive inferences. The mean MAPE of state estimation accuracy of V and θ is compared to see which factors in the conceptual model influence the estimation accuracy and in which way so that the research questions can be answered.

5.7.3 Analogic Inferences

The intended scope of generalization is the 33-bus power system where the Sample Mechanism is used to provide personalized privacy budgets to users. Therefore, the number of households per bus, threshold of the Sample Mechanism, privacy budgets, fraction of privacy conscious users, network size, day of the week, month of the year and state estimation techniques are varied to be able to make this generalization.

Chapter 6

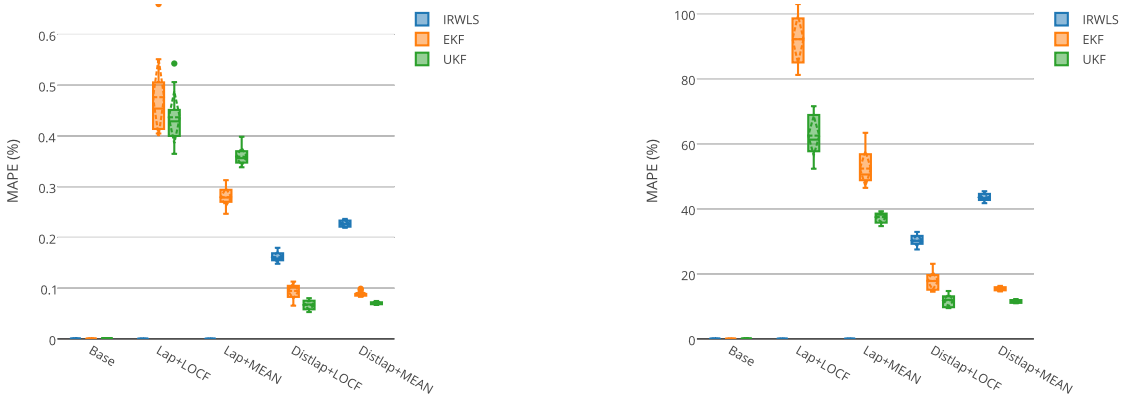
Results

In this chapter, the planned inferences described in the previous chapter are performed after the research was executed. All graphs represent the state estimation accuracy in MAPE. State estimation accuracy include both the accuracy of the estimation of voltage magnitude and voltage angle.

6.1 State Estimation and Imputation Technique

Results of varying the state estimation and imputation techniques can be seen in figures 6.1a and 6.1b. The IRWLS method did not converge for the untrusted setting where Laplace noise is added to each measurement z_i, so this result is not plotted. In the base scenario, where all measurements are available accurately and no differential privacy is guaranteed, the MAPE is approximately zero for all state estimators.

The Extended Kalman Filter is used in this thesis as state estimation technique, as it is more than 100 times faster than the Unscented Kalman Filter. In all scenarios, the IRWLS performs worse than the EKF and UKF. The EKF performs similar in the untrusted setting, probably due to the fact that the measurement covariance matrix R is not modeled optimally. The UKF performs better in the (partly) trusted scenario, as it is able to capture the non-linearity better and the measurement covariance matrix R is set to the right variances.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.1 Varying the different state estimators and imputation techniques in all scenarios.

The imputation technique MEAN performs slightly better than the LOCF. As LOCF is not compatible with the partly trusted scenario without compromising the differential privacy guarantee (as described in section 5.3), the MEAN imputation technique is used for the other experiments.

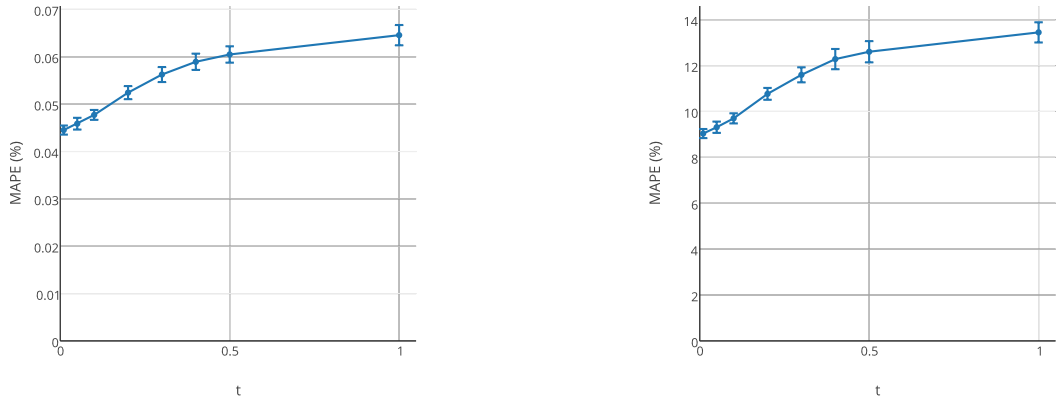
6.2 Differential Privacy

In this section, inferences related to differential privacy are made. Concepts that are elaborated are threshold, privacy budgets of the conservative group and moderate group, fraction of users in conservative group and composition size.

6.2.1 Threshold

The threshold t of the Sample Mechanism determines the privacy budget ε of the (distributed) Laplace Mechanism and which tuples are sampled. Note that the Sample Mechanism collapses to the Laplace Mechanism with t equal to 0.01 (lowest privacy budget).

In the (partly) trusted scenario, an increase in threshold will decrease the state estimation accuracy. The variance of the Laplace noise added to the aggregate is proportional to $\frac{1}{t^2}$. The probability of sampling tuples can be seen in figure 6.4, a higher threshold means that less measurements are sampled. As increasing the threshold decreases the state estimation accuracy, it is better to add more measurements than imputing more measurements with the mean in the (partly) trusted scenario.

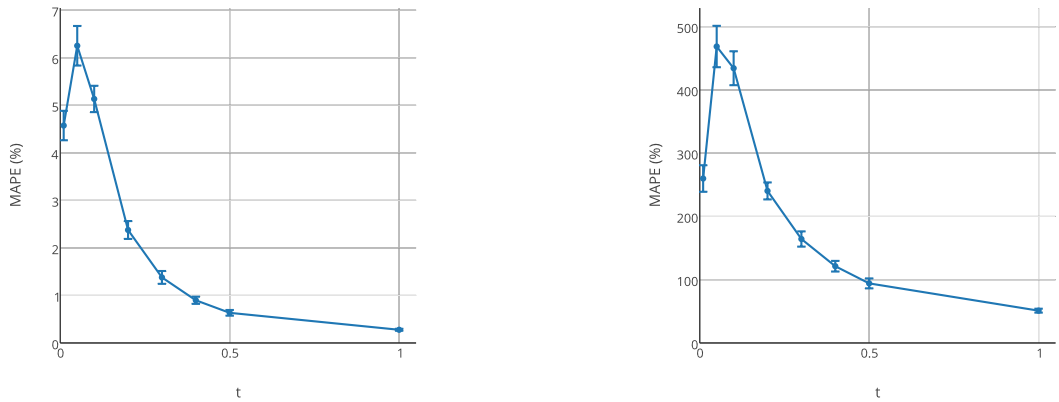


(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.2 Varying the threshold of the Sample Mechanism in the (partly) trusted scenario.

In the trusted scenario, an increase in threshold will increase the state estimation accuracy generally. This can be explained by imputing the missing data with the mean is more accurate than adding more measurements noised with Laplace. The variance of the noise added by the Laplace noise is $n \cdot 2(\frac{S(f)}{\varepsilon})^2$ (with $\varepsilon = t$ and n the number of samples) and is proportional to $\frac{n}{t^2}$. A higher measurement noise variance means that the true state is harder to track. The first spike can be explained by the fact that the measurement covariance matrix R is not optimal for $t > 0.01$, where less than 100% of the tuples are sampled.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.3 Varying the threshold of the Sample Mechanism in the untrusted scenario.

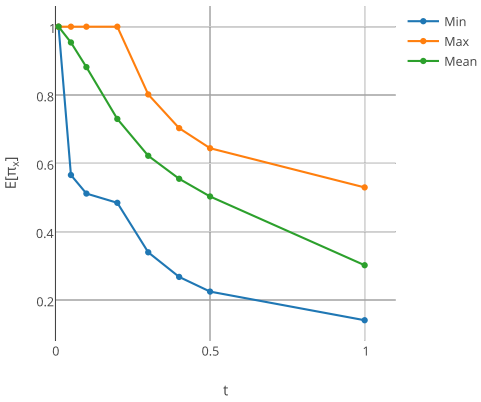


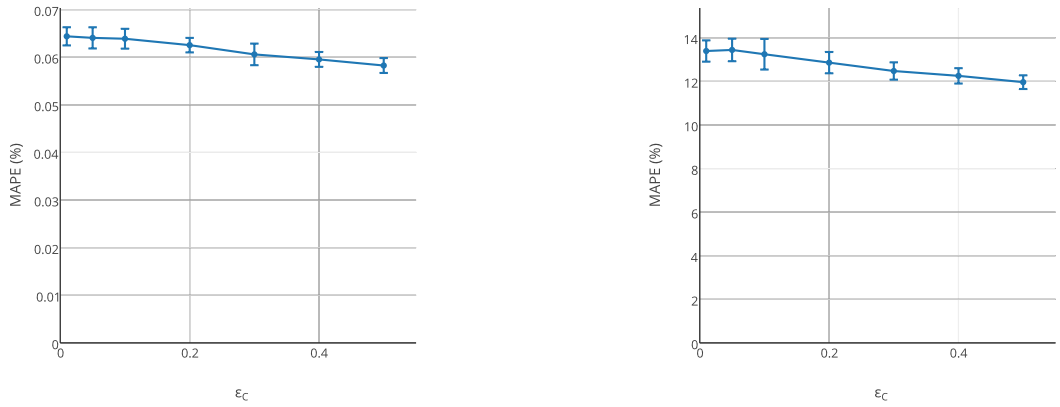
Fig. 6.4 Probability of sampling tuples by varying the threshold of the Sample Mechanism.

6.2.2 Privacy Budgets of Conservative Group

The privacy budget of users in the conservative group is selected at random between ε_C and ε_M, with ε_C the minimum. So if $\varepsilon_C \geq \varepsilon_M$, then the privacy budget is fixed at ε_C.

For both scenarios, the state estimation accuracy increases by increases the privacy budget of the conservative group. With a higher privacy budget, more measurements will be sampled (as seen in figure 6.7a) and thus the estimate of the load per bus and finally the accuracy will increase by a better imputation accuracy as described n the next paragraph. The accuracy increases faster than varying ε_M of the moderate group, as the probability of sampling tuples increases faster.

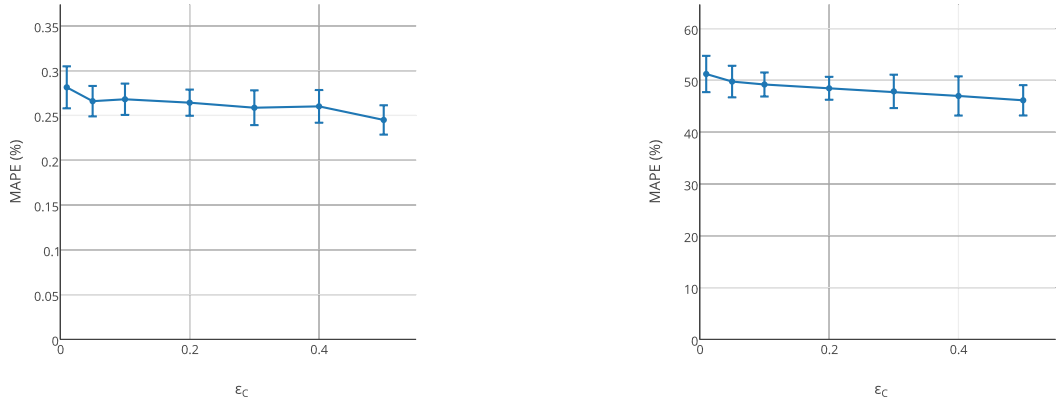
The imputation accuracy for the MEAN imputation technique by varying the probability of measurements sampled can be seen in figure 6.8. Here, the threshold is set to 1. The MEAN imputation technique performs worse if less measurements are sampled. This outcome is used for the next inferences involving differential privacy parameters.



(a) Voltage magnitude.

(b) Voltage angle.

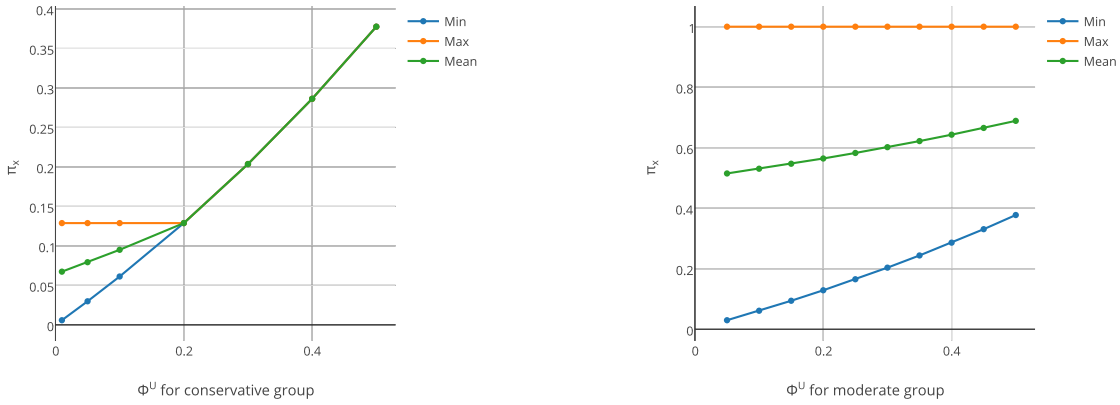
Fig. 6.5 Varying the privacy budget of conservative group in the (partly) trusted scenario.



(a) Voltage magnitude.

(b) Voltage angle.

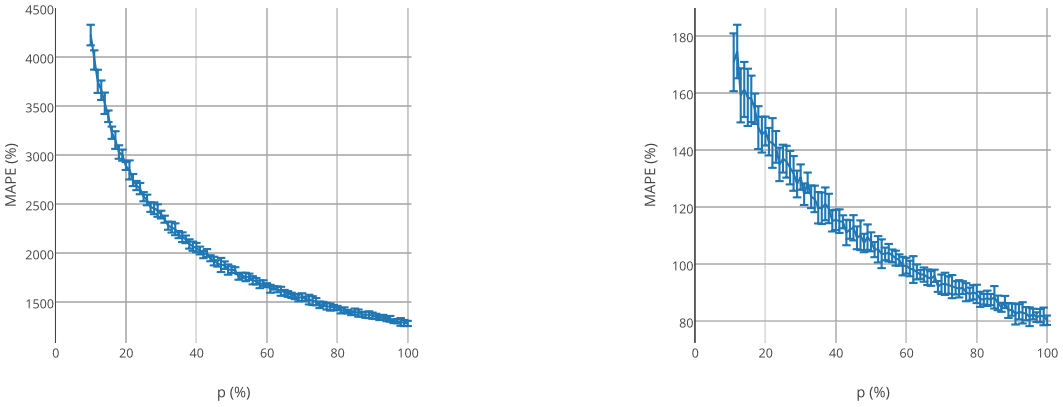
Fig. 6.6 Varying the privacy budget of conservative group in the untrusted scenario.



(a) Conservative group.

(b) Moderate group.

Fig. 6.7 Probability of sampling tuples by varying the privacy budgets of the conservative and moderate group.



(a) Untrusted scenario.

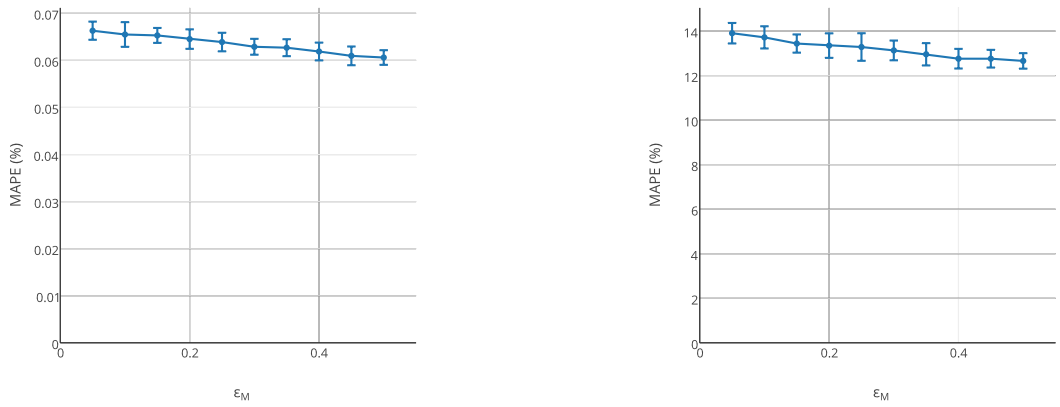
(b) (Partly) trusted scenario.

Fig. 6.8 Varying the probability of sampling measurements on the accuracy of the MEAN imputation technique.

6.2.3 Privacy Budgets of Moderate Group

The privacy budget of users in the conservative group is selected uniformly at random between ε_M and ε_L. For both scenarios, a higher privacy budget for users in the moderate group increases the state estimation accuracy. This is due to the same reason as increasing the

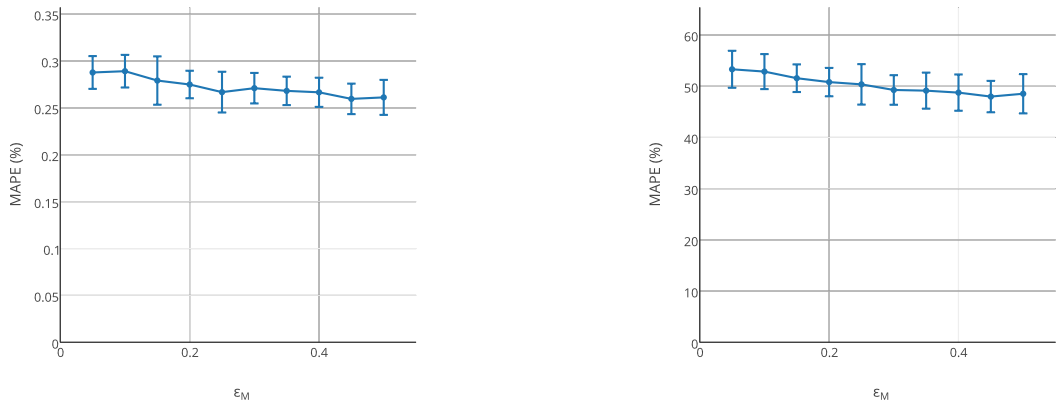
privacy budget of the conservative group described in the previous subsection. However, the state estimation accuracy increases slower by increasing the privacy budget of the moderate group than in the case of increasing the ε_C of the conservative group, due to the fact that the probability of sampling tuples increases more slowly (as seen in figure 6.7b). Also the conservative group is larger than the moderate group, assigning 54% and 37% of the users respectively.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.9 Varying the privacy budget of moderate group in the (partly) trusted scenario.



(a) Voltage magnitude.

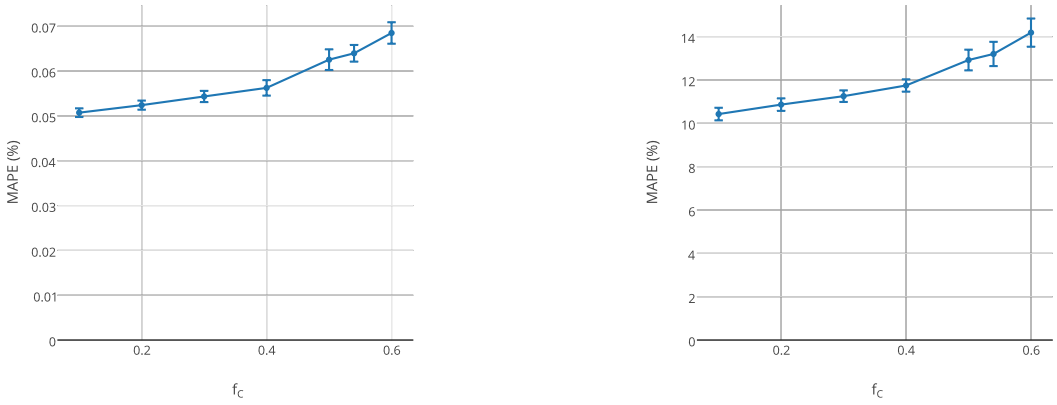
(b) Voltage angle.

Fig. 6.10 Varying the privacy budget of moderate group in untrusted scenario.

6.2.4 Fraction of Users in Conservative Group

The fraction of users in the conservative group f_C determines how many users are conservative. As the fraction of users in the moderate group stays constant, the fraction of users in the liberal group plus conservative group stays constant.

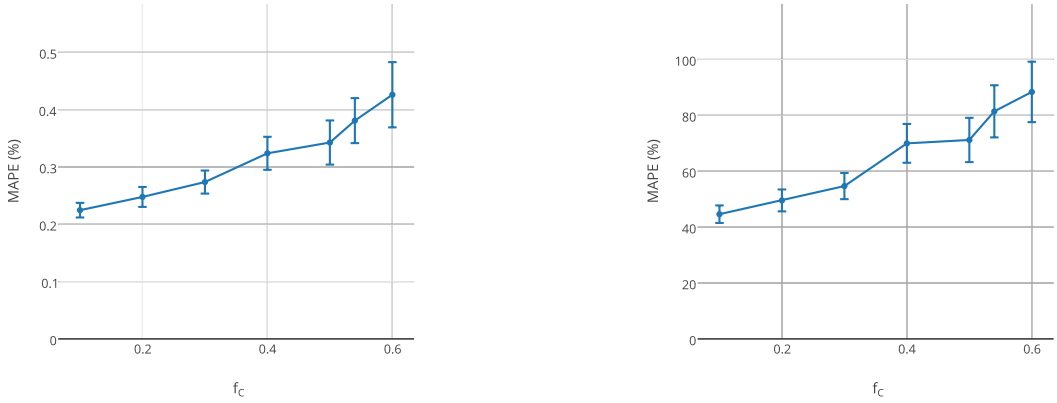
An increase in the fraction of users in the conservative group causes a decrease in state estimation accuracy. As more users are assigned to the conservative group, more users will get a lower privacy budget. A lower privacy budget means that less measurements are sampled and thus more measurements are imputed. This can be seen in figure 6.13a, where the mean of the probability that a measurement is sampled over all groups is plotted.



(a) Voltage magnitude.

(b) Voltage angle.

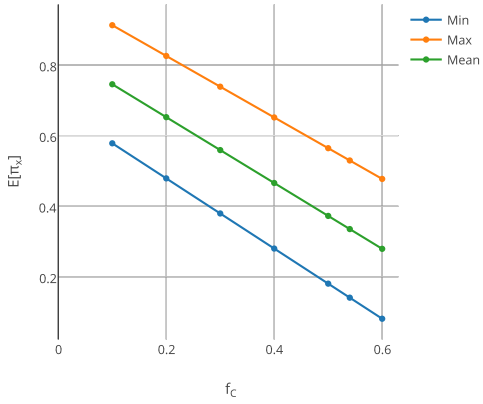
Fig. 6.11 Varying the fraction of users in conservative group in the (partly) trusted scenario.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.12 Varying the fraction of users in conservative group in the untrusted scenario.



(a) Conservative group.

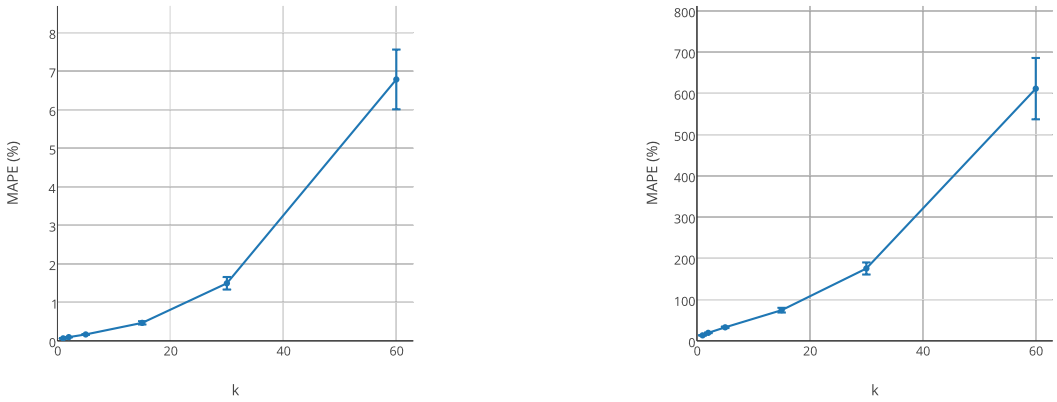
Fig. 6.13 Mean of probability of sampled tuples by varying the fraction of users in a group.

6.2.5 Composition Size

The composition size k is the number of sequential compositions the privacy budget needs to be guaranteed for. In both scenarios, the state estimation accuracy decreases fast with higher composition size k. In the untrusted scenario, only the values for composition sizes 1 and 2 are plotted. For higher composition sizes, the mean and standard deviation were too

high that indicates that the state estimator could not track the state anymore. In sequential composition, the privacy budget that needs to be guaranteed deteriorates with $\frac{1}{k}$. This means that the variance of the noise that is added is proportional to k^2. For the (partly) trusted scenario, the standard deviation of a composition size of 60 indicates that the EKF was not able to track the true state.

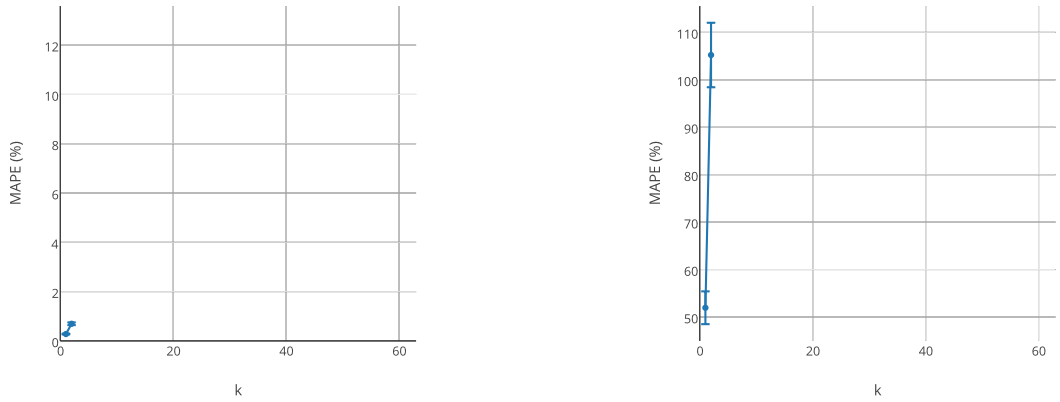
As the privacy budget and threshold deteriorates with $\frac{1}{k}$, the number of sampled tuples will slightly increase according to definition 5. This can be seen in figures 6.16a and 6.16b for respectively the default values $\varepsilon_C = 0.01$ and $\varepsilon_M = 0.2$, for the liberal group all tuples are sampled if $t = \varepsilon_L$. The probability that tuples are sampled for users in the moderate group increases faster with a higher k than the conservative group. In both groups the values converge to $\lim_{k\to\infty} = \frac{e^{(\phi^x \mathscr{U})/k}-1}{e^{(t/k)}-1} = \frac{\phi^x \mathscr{U}}{t}$ and thus the effects that increases the estimation accuracy damp with a higher k (while the variance of the Laplace noise increases with k^2). Please note that the privacy budget here is constant, in contradiction with the privacy budgets uniformly drawn for users in the state estimation accuracy experiments.



(a) Voltage magnitude.

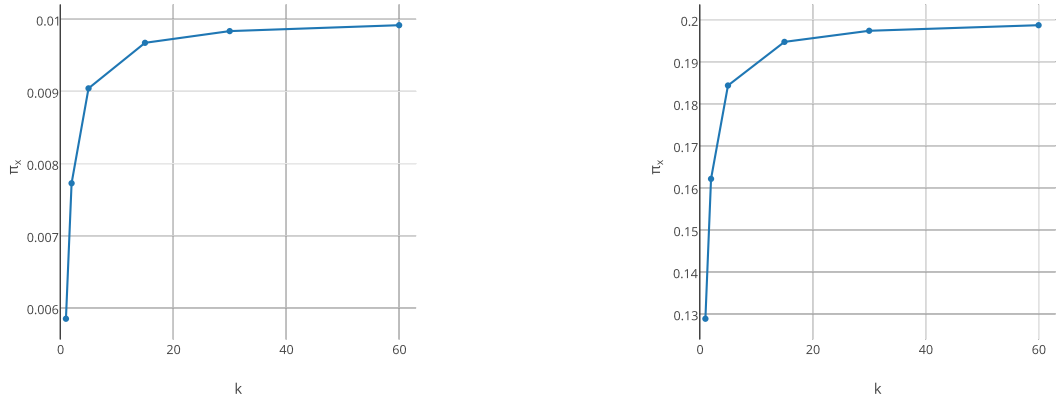
(b) Voltage angle.

Fig. 6.14 Varying the composition size in the (partly) trusted scenario.



(a) Voltage magnitude. (b) Voltage angle.

Fig. 6.15 Varying the composition size in the untrusted scenario.



(a) For ε_C. (b) For ε_M.

Fig. 6.16 Probability of sampling tuples by varying the composition size k.

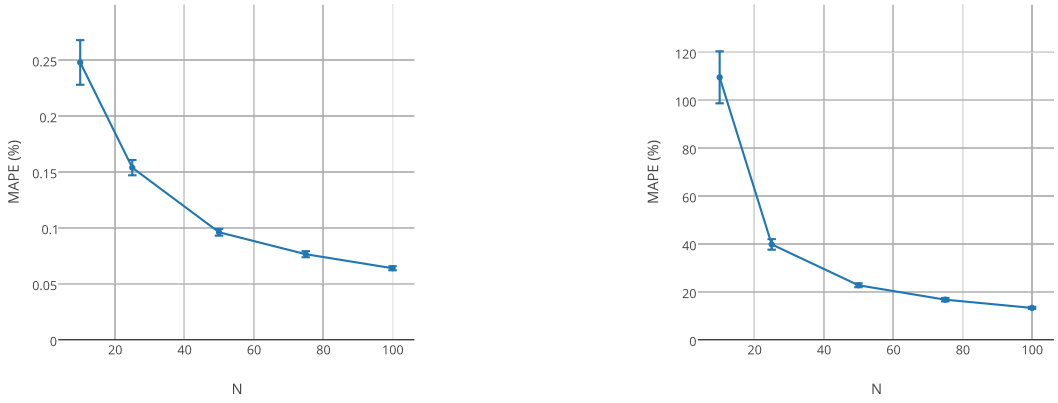
6.3 Context

In this section, the context of the treatment is varied. The components of the context that are examined are number of households per substation that participated and dataset characteristics.

6.3.1 Number of Households per Substation

The number of households participating in a LV area (denoted as N) is the number of households connected to a substation. In both scenarios, the MAPE decreases with more participating smart meters per substation. This is mostly due to the scaling of loads per bus to the maximum load of the bus, that causes a decrease in the variance of the load and state changes. This can be seen in figures 6.19a and 6.19b, where the change in real power has more influence on changes in the state due to the higher values. Lower changes in the state result in higher state estimation accuracy, especially in this setting where process noise matrix Q is set to 10% of the max state change.

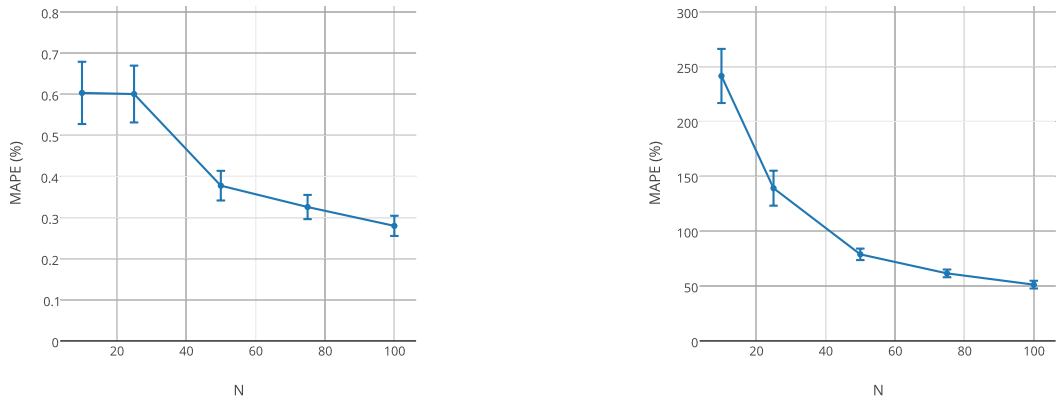
For the (partly) trusted scenario, the Laplace noise added is not changed as it is dependent on the threshold t, composition size k and the sensitivity $S(f)$. For the untrusted scenario, the variance of the sum of loads by adding Laplace noise is at most $2(\frac{S(f)}{k\varepsilon})^2 \cdot N$ (as described in section 5.3). This should slightly damp the effect of the lower state changes due to scaling the dataset to the maximum load of the bus.



(a) Voltage magnitude.

(b) Voltage angle.

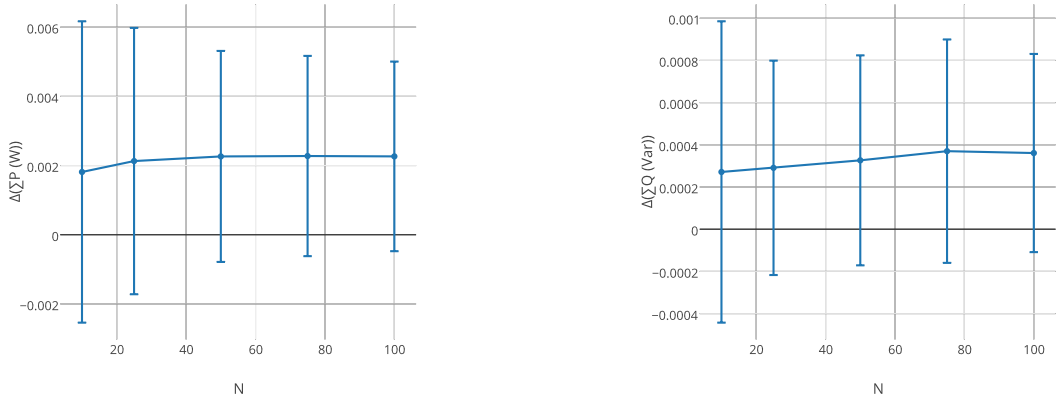
Fig. 6.17 Varying the number of households in the (partly) trusted scenario.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.18 Varying the number of households in the untrusted scenario.



(a) Real power.

(b) Reactive power.

Fig. 6.19 Load change of bus 2 by varying the number of households (without noise).

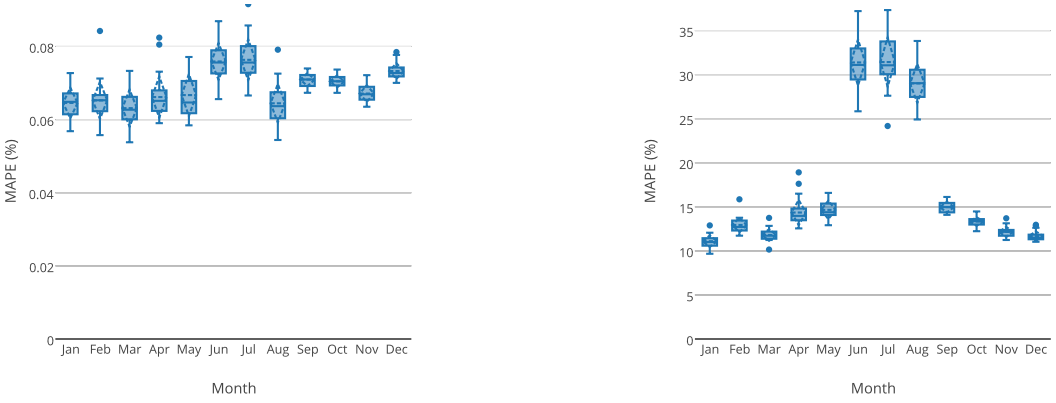
6.3.2 Dataset Characteristics

The dataset is generated using the month of the day and type of day (weekday or weekend). Effects of varying these parameters for generating the dataset on the state estimation accuracy is described in this section.

The state estimation accuracy for the months Januari, February, March, April, October, November and December are approximately equal. The accuracy of the voltage angle is significant higher for the months June, July and August and slightly higher for May and September. This is due to the dataset characteristics of these months. In figures A.1a, A.1b, A.2a and A.2b, it can be seen that the standard deviation of the loads in these months are lower (as well as the mean). This is due to the lower temperature in the UK in June, July and August that is modeled in the CREST dataset. Temperature modifier values are 1.63, 1.821, 1.595, 0.867, 0.763, 0.191, 0.156, 0.087, 0.399, 0.936, 1.561 and 1.994 for the months January till December. The activity probability of heating appliances are directly related to the temperature modifier.

As the loads are scaled (as described in section 5.1) and the maximum loads of the months are approximate equal, the standard deviation of the months June, July and August will be scaled with a higher factor than other months. This results in a higher state changes, which has impact on the state estimation accuracy. The higher state changes can be seen in figures A.3b, A.3a, A.4b and A.4a.

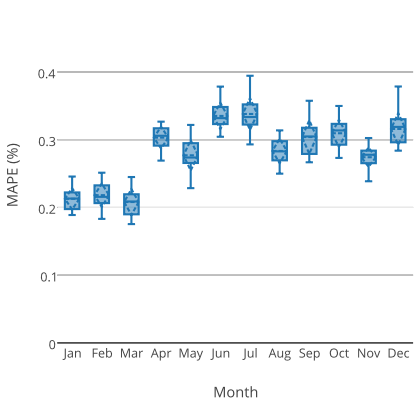
Weekday



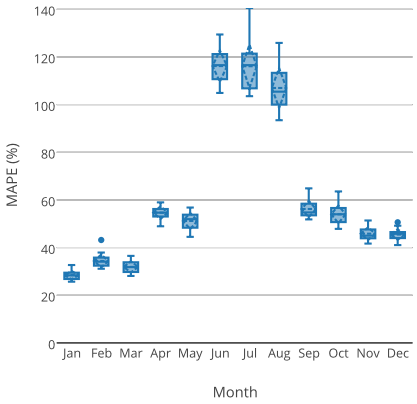
(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.20 Varying the month of the year during weekdays in the (partly) trusted scenario.



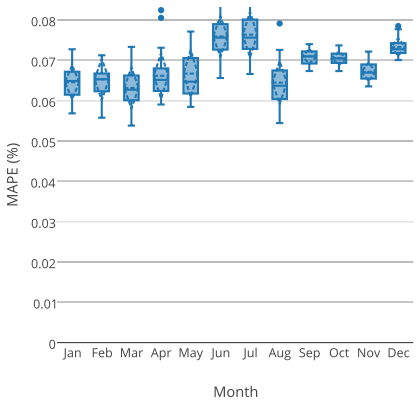
(a) Voltage magnitude.



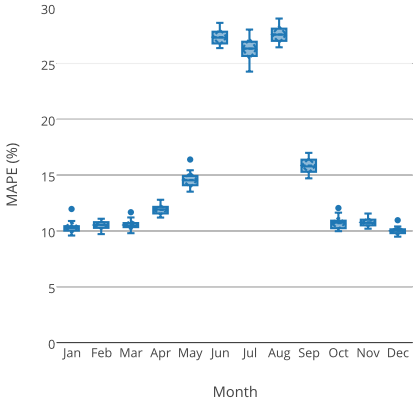
(b) Voltage angle.

Fig. 6.21 Varying the month of the year during weekdays in the untrusted scenario.

Weekend

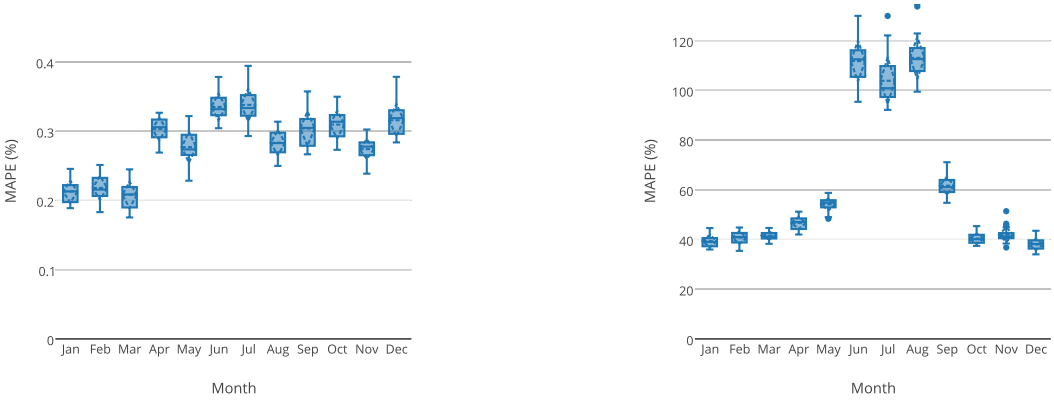


(a) Voltage magnitude.



(b) Voltage angle.

Fig. 6.22 Varying the month of the year during weekend in the (partly) trusted scenario.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. 6.23 Varying the month of the year during weekend in the untrusted scenario.

6.4 Reflection on Research Questions

In the previous subsections, descriptive and abductive inferences are made. In this section, the descriptive inferences are linked to the research questions.

Effect questions: ask what effect an artifact in a context has.

1. What is the effect of the Sample Mechanism on state estimation accuracy?

   In the (partly) trusted scenario, the sample mechanism provides less state estimation accuracy with a higher threshold. Imputing missing measurements with the mean is less accurate than adding more measurements with a higher variance of the Laplace noise on the aggregated value in this scenario. In the untrusted scenario, the Sample Mechanism increases the state estimation accuracy with a higher threshold. The variance added to each measurement grows quadratic with a higher threshold and therefore the effect sampling more measurements is small. This implies that the sample mechanism is effective in the untrusted scenario, but ineffective in the (partly) trusted scenario.

Trade-off questions: ask what the difference between effects of different artifacts in the same context are.

2. What is the effect of different state estimation and imputation techniques on state estimation accuracy?

Three state estimators are evaluated: the Iterative Recursive Weighted Least Squares (IRWLS), Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF). The two imputation techniques are last observation carried forward (LOCF) and mean (MEAN). While the IRWLS did not converge for the untrusted scenario, the it performed worse than the EKF and UKF. The UKF performed slightly better in the (partly) trusted scenario and about the same for the untrusted scenario. The MEAN imputation technique performed better than LOCF in both scenarios.

3. What is the effect of varying personalized privacy budgets on state estimation accuracy?

   The privacy budgets of the conservative group and moderate group are varied. In both scenarios, increasing both budgets increases the state estimation accuracy. For the conservative group, the state estimation accuracy increases faster than changing the budget of the moderate group. This is due to the fact that more measurements are sampled in the conservative group by increasing the budget of the conservative group and that the conservative group is larger than the moderate group.

   An increase in the fraction of users in the conservative group decreases the state estimation accuracy. This is due to the low privacy budgets of users of the conservative group, that is chosen uniformly at random between privacy budget of the conservative group and the moderate group. Users of the liberal group become conservative in this experiment.

4. What is the effect of guaranteeing the privacy budgets under composition on state estimation accuracy?

   Sequential composition is chosen to model the composition of measurements over time. The default privacy budgets for all groups vary from 0.1 to 1.0. This means that the privacy budgets vary from $\frac{0.1}{k}$ and $\frac{1.0}{k}$ under k-fold composition. As the variance of the Laplace noise added to both scenarios is proportional to k^2, the state estimation accuracy deteriorates highly with a higher composition size. The EKF was not able to track the real state in the untrusted scenario for $k \geq 5$.

Sensitivity questions: ask what happens if the context is changed.

5. What is the effect of the number of participating smart meters in low-voltage areas on state estimation accuracy?

   In both scenarios, an increase in the number of participating smart meters per bus increases the state estimation accuracy. This is mainly due to scaling the dataset to the maximum load per bus, so that the variance of the state change decreases with more participating smart meters.

6. What is the effect of changing the day of the week and month of the year on state estimation accuracy?

   The state estimation accuracy is approximately the same for all months for weekend and weekdays, except the months May, June, July, August and September. In these months, the dataset generator uses a temperature modifier such that heat appliances turned on less often. As these appliances causes spikes in the loads, this means that the standard deviation of the mean state change is lower and thus the accuracy of the state estimation increases.

Chapter 7

Conclusion

In this chapter, implications for the context are given. Obtained knowledge is discussed and a brief reflection on the research questions is elaborated. Also a discussion of the results and future research directions are given.

7.1 Conclusion

In this thesis, the Sample Mechanism is used to give users a personalized privacy budget. The Sample Mechanism samples tuples (measurements) from users and then utilizes a differentially private mechanism so that the two sources of randomness can guarantee the user's differential privacy budget. Sampling is seen as intermittent transmission of measurements to the aggregator in this thesis. The Sample Mechanism depends on the threshold parameter, which balances between the probability of sampling measurements and the noise added by the differentially private mechanism. The main research question was about exploring the effects of using the Sample Mechanism on state estimation accuracy. Single-case experiments in the form of (Monte Carlo) simulations are performed to obtain results.

Four scenarios are tested, the first scenario is the base scenario where no measurements are noised. In the untrusted scenario, each user is guaranteed ε-differential privacy without any trust in the aggregator. In the fully trusted scenario, the aggregator receives noise-free measurements of the users and guarantee them ε-differential privacy for the users. In the partly trusted scenario, the user transmits their measurement plus Gamma noise. The sum of these Gamma noises result in the Laplace Mechanism. By using homomorphic encryption, these non-private measurements can not be obtained easily by the aggregator. The fully and partly trusted scenarios are combined as the resulting Laplace Mechanism is equal and is called "(partly) trusted scenario".

A non-linear model of a power system is used. To estimate the state of this system, three state estimators are tested. The Extended Kalman Filter (EKF) is used due to its moderate state estimation accuracy and computation speed. The UKF performed better and the IRWLS performed worse in terms of accuracy. To impute the missing measurements due to the Sample Mechanism, the mean imputation technique is chosen as it performed better than the last-observation-carried-forward (LOCF) technique.

The Sample Mechanism is effective for the untrusted scenario, while it is ineffective for the (partly) trusted scenario. This is measured by varying the threshold of the Sample Mechanism. If the threshold is equal to the minimum personalized privacy budget, the Sample Mechanism simplifies to the Laplace Mechanism. For the untrusted scenario, it can be seen that a higher threshold increases the state estimation accuracy. For the (partly) trusted scenario, the accuracy decreases with a higher threshold. This means that the Sample mechanism can be used in the untrusted setting to gain significant accuracy. In the (partly) trusted scenario, the accuracy will decrease slightly. There is a trade-off between the probability of not transmitting measurements that results in imputing the missing measurements with the mean and the noise added by the Laplace Mechanism.

Increasing the privacy budgets of the conservative and moderate groups result in a higher state estimation accuracy. An increase in the fraction of users in the conservative group decreases the state estimation accuracy. The state estimation accuracy under k-fold sequential composition is significant higher for $k > 1$. For the untrusted scenario, the state estimator could not track the true state accurately for $k \geq 5$.

The number of participative smart meters per bus has great influence on the state estimation accuracy. In both scenarios, the accuracy increases with a higher number of participants. In the months May, June, July, August and September the state estimation accuracy decreases due to the higher temperature modifier in the dataset generator and scaling of the dataset to the maximum load of the bus. This leads to a higher variance in the state changes and thus a lower state estimation accuracy.

7.2 Discussion and Future Research

This thesis generalizes for the 33-bus power network with a single generator. These generalizations could be extended to more types of power networks by performing similar inferences in another contexts. It could even be further generalized to for example water networks or vehicle-to-vehicle communication.

The Extended Kalman Filter (EKF) is used as state estimator. In theory, the particle filter will outperform both state estimators and thus better results can be achieved. Downside of

using a slower state estimator is that the measurements become less valuable as can be read in the next paragraph. Also the noise covariance matrix R is set to N times the variance of the Laplace noise added in the untrusted setting with the mean imputation method, which does not approximate the covariance matrix optimally. This can be improved if knowledge about privacy preferences is obtained by the network operator. This can be done by making inferences about when users do not send their measurements, however this is sensitive for measurements lost due to other variables. The privacy preferences could perhaps also be send to the network operator automatically (as it is tunable by the user).

Delays such as transmitting measurements, decryption and performing dynamic state estimation have not been modeled in this thesis. However delays and unsynchronized measurements in distributed state estimation have a huge impact on state estimation accuracy [51]. This aspect of distributed state estimation is not neglectable and should be further researched in the context of this thesis.

A maximum deviation of 0.7% for the voltage magnitude and 0.7 crad ($0.7 \cdot 10^{-2}$ rad) for the voltage angle is considered acceptable [33]. Most settings researched in this thesis could provide a maximum deviation of 0.7% for the voltage magnitude. However, for the voltage angle this can not be inferred from the collected MAPE performance index. Also, the performance index in [33] uses a compliance probability of 95%. To see whether the treatment in this thesis is useful in a real-world setting, requirements need to be operationalized by using proper units. Another performance indicator could be the maximum of the MAPE, as the acceptable deviation in [33] has a maximum deviation.

The synthetic dataset that is used for power loads of households is based on consumption data of Irish households [48]. Factors as climate, socioeconomic status and number of occupants influence the power consumption [31]. Therefore, the performance of state estimation can vary among locations worldwide. Also, the electricity demand has been normalized to not exceed the maximum demand of each bus of the 33-bus network. In reality, the number of households per bus is related to the maximum load of each household.

The homomorphic encryption scheme used in the partly trusted scenario can be attacked such that differential privacy can not be guaranteed anymore [1]. The aggregator can lie about the cluster size and non-responding nodes. This thesis did not implement the homomorphic encryption scheme. To fully understand aggregating measurements this way, more research is needed in the context of the Sampling Mechanism.

The last observation carried forward imputation technique is incompatible with the partly trusted scenario described in this thesis, as the additive homomorphic encryption scheme used in the partly trusted scenario does not allow to recover individual measurements. However, it could be possible to make this technique compatible by using buffers [7]. These buffers can

contain the last observation or a prediction of electricity consumption. Note that a prediction would mean that additional information is given to the aggregator which has implications for differential privacy.

Power utility companies need to invoice their customers. They do this by receiving non-private power consumption data of the customer at time intervals of one month, six months or one year or (near) real-time in a smart grid setting [37]. This means that the privacy budget of the user can be exceeded by possibly providing additional measurements. A model that could overcome this issue and works with both aggregation and billing is provided in [10].

This thesis uses pure differential privacy to be guaranteed to users. However, relaxations are developed that provide higher accuracy without significant lower privacy guarantees [17]. For example approximate differential privacy could be used to achieve better state estimation accuracy. Also, group privacy has not been included in this thesis and is an interesting as a topic for further research.

References

[1] Ács, G. and Castelluccia, C. (2011). *I Have a DREAM! (DiffeRentially privatE smArt Metering)*, pages 118–132. Springer Berlin Heidelberg, Berlin, Heidelberg.

[2] Al-wakeel, A. (2016). *Integrated Load and State Estimation Using Domestic Smart Meter*. PhD thesis, Cardiff University.

[3] Al-Wakeel, A., Wu, J., and Jenkins, N. (2016). State estimation of medium voltage distribution networks using smart meter measurements. *Applied Energy*, 184:207 – 218.

[4] Anderson, B., Lin, S., Newing, A., Bahaj, A., and James, P. (2016). Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. *Computers, Environment and Urban Systems*, pages –.

[5] Baran, M. E. and Wu, F. F. (1989). Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Transactions on Power Delivery*, 4(2):1401–1407.

[6] Barbosa, P., Brito, A., and Almeida, H. (2016). A technique to provide differential privacy for appliance usage in smart metering. *Information Sciences*, 370–371:355 – 367.

[7] Bemporad, A. (1998). Predictive control of teleoperated constrained systems with unbounded communication delays. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No.98CH36171)*, volume 2, pages 2133–2138 vol.2.

[8] Bhattacharyya, S., Wang, Z., Cobben, J., and Kling, W. (2008). Analysis of power quality performance of the dutch medium and low voltage grids.

[9] Bhela, S., Kekatos, V., and Veeramachaneni, S. (2016). Enhancing Observability in Distribution Grids using Smart Meter Data. *ArXiv e-prints*.

[10] Borges, F., Demirel, D., Böck, L., Buchmann, J., and Mühlhäuser, M. (2014). A privacy-enhancing protocol that provides in-network data aggregation and verifiable smart meter billing. In *2014 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6.

[11] Bun, M. and Steinke, T. (2016). *Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds*, pages 635–658. Springer Berlin Heidelberg, Berlin, Heidelberg.

[12] Charles, A. S., Balavoine, A., and Rozell, C. J. (2016). Dynamic filtering of time-varying sparse signals via ℓ_1 minimization. *IEEE Transactions on Signal Processing*, 64(21):5644–5656.

[13] Chou, J.-S., Hsu, Y.-C., and Lin, L.-T. (2014). Smart meter monitoring and data mining techniques for predicting refrigeration system performance. *Expert Systems with Applications*, 41(5):2144 – 2156.

[14] Darby, S. (2010). Smart metering: what potential for householder engagement? *Building Research & Information*, 38(5):442–457.

[15] Deng, M., Fan, Z., Liu, Q., and Gong, J. (2016). A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets. *International Journal of Geo-Information*, 13(5).

[16] Dwork, C. (2006). Differential privacy.

[17] Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

[18] Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *CoRR*, abs/1603.01887.

[19] Dwork, C., Rothblum, G. N., and Vadhan, S. (2010). Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60.

[20] Ebadi, H., Sands, D., and Schneider, G. (2015). Differential privacy: Now it's getting personal. In *Proceedings of the 42Nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '15, pages 69–81, New York, NY, USA. ACM.

[21] Eibl, G. and Engel, D. (2016). Differential privacy for real smart metering data. *Computer Science - Research and Development*, pages 1–10.

[22] El-hawary, M. E. (2014). The smart grid—state-of-the-art and future trends. *Electric Power Components and Systems*, 42(3-4):239–250.

[23] Emami, K., Fernando, T., Iu, H. H. C., Trinh, H., and Wong, K. P. (2015). Particle filter approach to dynamic state estimation of generators in power systems. *IEEE Transactions on Power Systems*, 30(5):2665–2675.

[24] Farokhi, F., Milosevic, J., and Sandberg, H. (2016). Optimal state estimation with measurements corrupted by laplace noise.

[25] Geng, Q., Kairouz, P., Oh, S., and Viswanath, P. (2015). The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184.

[26] Ghahremani, E. and Kamwa, I. (2011). Dynamic state estimation in power system by applying the extended kalman filter with unknown inputs to phasor measurements. *IEEE Transactions on Power Systems*, 26(4):2556–2566.

[27] Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing. *The American Statistician*, 61(1):79–90. PMID: 17401454.

[28] Huang, Y. F., Werner, S., Huang, J., Kashyap, N., and Gupta, V. (2012). State estimation in electric power grids: Meeting new challenges presented by the requirements of the future grid. *IEEE Signal Processing Magazine*, 29(5):33–43.

[29] Huanyuan, C., Xindong, L., Caiqi, S., and Cheng, Y. (2011). Power system dynamic state estimation based on a new particle filter. *Procedia Environmental Sciences*, 11:655 – 661.

[30] Jorgensen, Z., Yu, T., and Cormode, G. (2015). Conservative or liberal? personalized differential privacy. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1023–1034.

[31] Kavousian, A., Rajagopal, R., and Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, 55:184 – 194.

[32] Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. (2014). Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392.

[33] Liu, J., Tang, J., Ponci, F., Monti, A., Muscas, C., and Pegoraro, P. A. (2012). Trade-offs in pmu deployment for state estimation in active distribution grids. *IEEE Transactions on Smart Grid*, 3(2):915–924.

[34] Lu, R. (2016a). *Privacy-Enhancing Aggregation Techniques for Smart Grid Communications*. Wireless Networks. Springer International Publishing.

[35] Lu, R. (2016b). *Privacy-Enhancing Aggregation Techniques for Smart Grid Communications*. Springer International Publishing.

[36] Lu, R., Liang, X., Li, X., Lin, X., and Shen, X. (2012). Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications. *IEEE Transactions on Parallel and Distributed Systems*, 23(9):1621–1631.

[37] McKenna, E., Richardson, I., and Thomson, M. (2012). Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, 41:807 – 814. Modeling Transport (Energy) Demand and Policies.

[38] McSherry, F. D. (2009). Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 19–30, New York, NY, USA. ACM.

[39] Mohammed, N., Chen, R., Fung, B. C., and Yu, P. S. (2011). Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 493–501, New York, NY, USA. ACM.

[40] Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., and Irwin, D. (2010). Private memoirs of a smart meter. In *Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, BuildSys '10, pages 61–66, New York, NY, USA. ACM.

[41] Nijhuis, M., Gibescru, M., and Cobben, S. (2015). Clustering of low voltage feeders from a network planning perspective.

[42] Nissenbaum, H. (2010). *Privacy in Context: Technologyt, Policy, and the Integrity of Social Life*. Stanford Law Books.

[43] Nor, N. M., Jegatheesan, R., and Nallagowden, I. P. (2008). Newton-raphson state estimation solution employing systematically constructed jacobian matrix. *International Journal of Electrical*, 2(6).

[44] Oh, S. and Viswanath, P. (2013). The composition theorem for differential privacy. *CoRR*, abs/1311.0776.

[45] Peppanen, J., Zhang, X., Grijalva, S., and Reno, M. J. (2016). Handling bad or missing smart meter data through advanced data imputation. In *2016 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5.

[46] Pietaloka, D. (2015). Distributed-dynamic state estimation with scada and pmu. Master's thesis, TU Delft.

[47] Qi, J., Sun, K., Wang, J., and Liu, H. (2016). Dynamic state estimation for multi-machine power system by unscented kalman filter with enhanced numerical stability. *IEEE Transactions on Smart Grid*, PP(99):1–1.

[48] Richardson, I., Thomson, M., Infield, D., and Clifford, C. (2010). Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10):1878 – 1887.

[49] Rottondi, C. E. M. (2013). *Privacy Preserving Data Collection in The Automatic Metering Infrastructure of Smart Grids*. PhD thesis, Politechno Di Milano.

[50] Saikia, Anupam and Mehta, Ram Krishna (2016). Power system static state estimation using kalman filter algorithm. *Int. J. Simul. Multisci. Des. Optim.*, 7:A7.

[51] Samarakoon, K., Wu, J., Ekanayake, J., and Jenkins, N. (2011). Use of delayed smart meter measurements for distribution state estimation. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–6.

[52] Sandberg, H., Dán, G., and Thobaben, R. (2015). Differentially private state estimation in distribution networks with smart meters. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4492–4498.

[53] Shiryaev, A. N. (1996). *Probability*. Springer New York.

[54] Sinopoli, B., Schenato, L., Franceschetti, M., Poolla, K., Jordan, M. I., and Sastry, S. S. (2004). Kalman filtering with intermittent observations. *IEEE Transactions on Automatic Control*, 49(9):1453–1464.

[55] Smith, M. M., Powell, R. S., Irving, M. R., and Sterling, M. J. H. (1991). Robust algorithm for state estimation in electrical networks. *IEE Proceedings C - Generation, Transmission and Distribution*, 138(4):283–288.

[56] Upadhyay, U., Rathore, S., Vashnav, G., and Khandelwal, A. (2016). Recent development in power system dynamic state estimation. *SSRG International Journal of Electrical and Electronics Engineering*, 3.

[57] U.S. Department of Energy (2009). Smart grid system report.

[58] Valverde, G. and Terzija, V. (2011). Unscented kalman filter for power system dynamic state estimation. *IET Generation, Transmission Distribution*, 5(1):29–37.

[59] Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.

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

[61] Xygkis, T. C., Karlis, G. D., Siderakis, I. K., and Korres, G. N. (2014). Use of near real-time and delayed smart meter data for distribution system load and state estimation. In *MedPower 2014*, pages 1–6.

[62] Yu, C. N., Mirowski, P., and Ho, T. K. (2016). A sparse coding approach to household electricity demand forecasting in smart grids. *IEEE Transactions on Smart Grid*, PP(99):1–11.

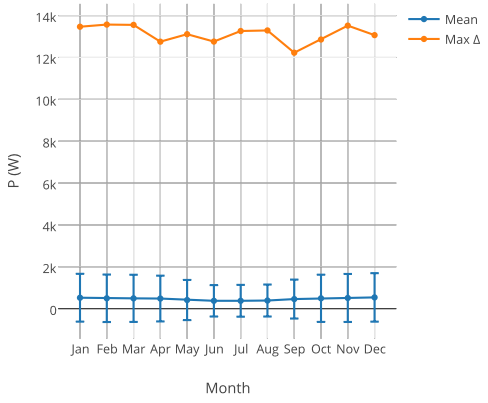
[63] Zoha, A., Gluhak, A., Imran, M. A., and Rajasegarar, S. (2012). Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors*, 12(12):16838–16866.

Appendix A

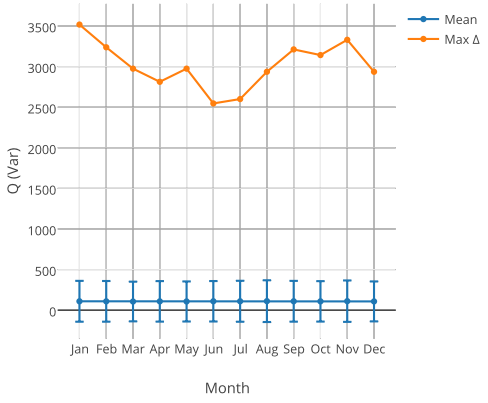
Dataset Characteristics

The dataset comprises of 3000 load profiles. The dataset can be downloaded on https://github.com/rutgerprins/thesis. Some characteristics of the dataset can be found below. WD stands for weekday and WE stands for a day in the weekend. Please note that these figures are scaled to the maximum load of each bus as described in chapter 5.

The mean, standard deviation and maximum change per month with $N = 100$ smart meters per bus are plotted in the figures A.1a, A.1b, A.2a and A.2b.

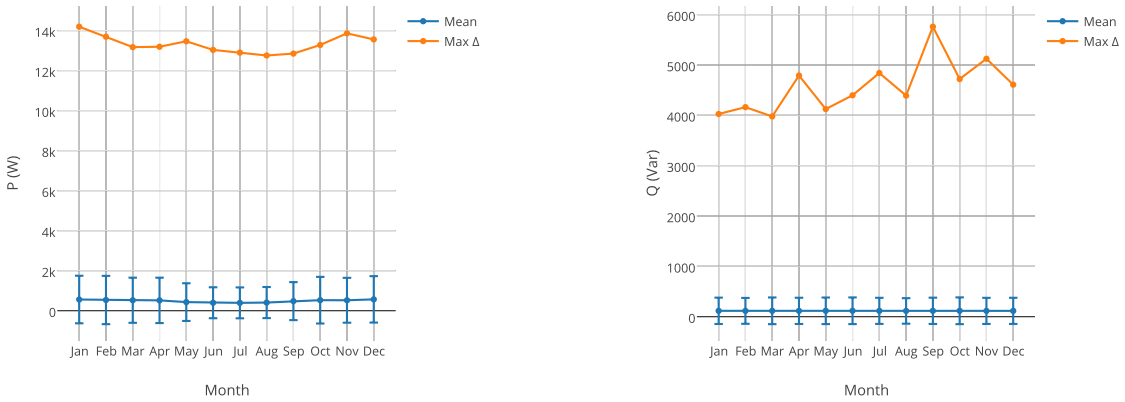


(a) Real power.



(b) Reactive power.

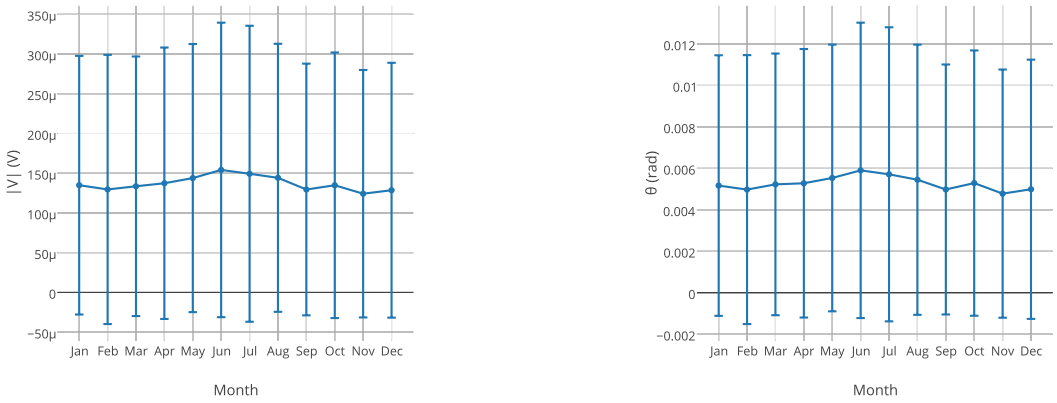
Fig. A.1 Loads during Weekday.



(a) Real power. (b) Reactive power.

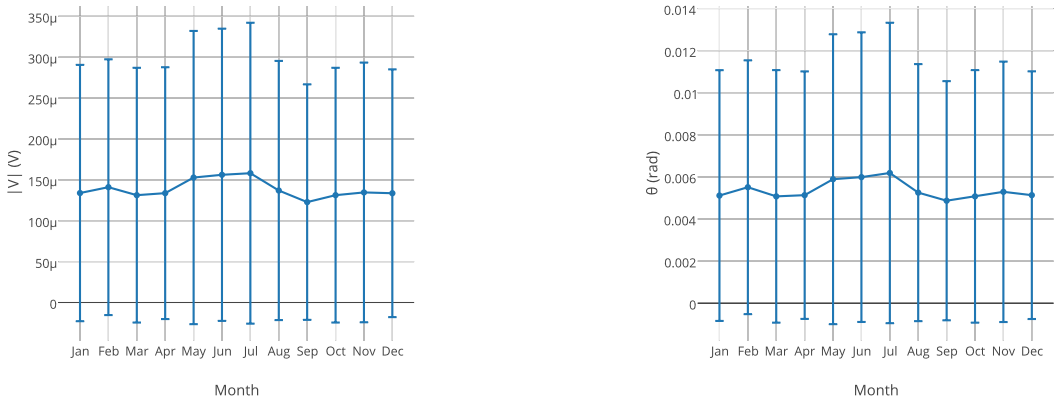
Fig. A.2 Loads during Weekend.

The mean and standard deviation of the maximum state change per month with $N = 100$ smart meters per bus are plotted in the figures A.3b, A.3a, A.4b and A.4a.



(a) Voltage magnitude. (b) Voltage angle.

Fig. A.3 State Change during Weekday.



(a) Voltage magnitude.

(b) Voltage angle.

Fig. A.4 State Change during Weekend.

The maximum load for individual and per bus (for each N) can be seen in table A.1.

Table A.1 Summary statistics of load profiles for May WD

Number of load profiles:	10	25	50	100
Max P (W)	14308	15719	15719	17060
Max Q (var)	4124	4124	4124	4124
Max $\sum P$ per substation (W)	23750	40067	70324	110065
Max $\sum Q$ per substation (var)	10007	15983	18659	27609

The aggregated active and reactive power for each dataset can be seen in figures A.5 and A.6. Individual load of 6 random chosen profiles for bus 2 can be seen in figure A.7.

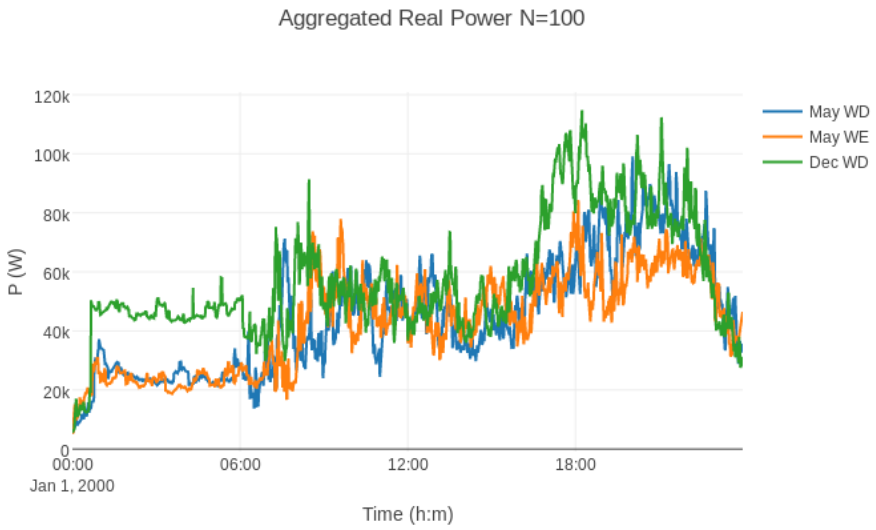


Fig. A.5 Aggregate of all 3000 load profiles.

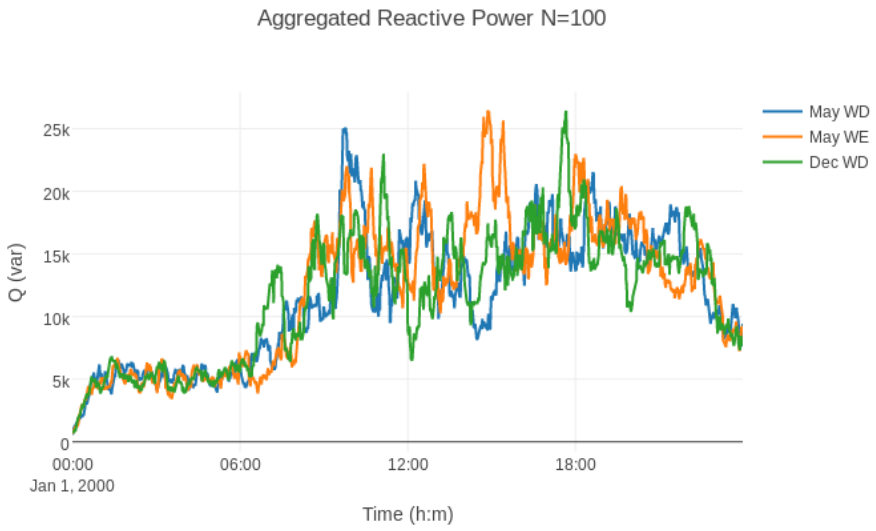


Fig. A.6 Aggregate of all 3000 load profiles.

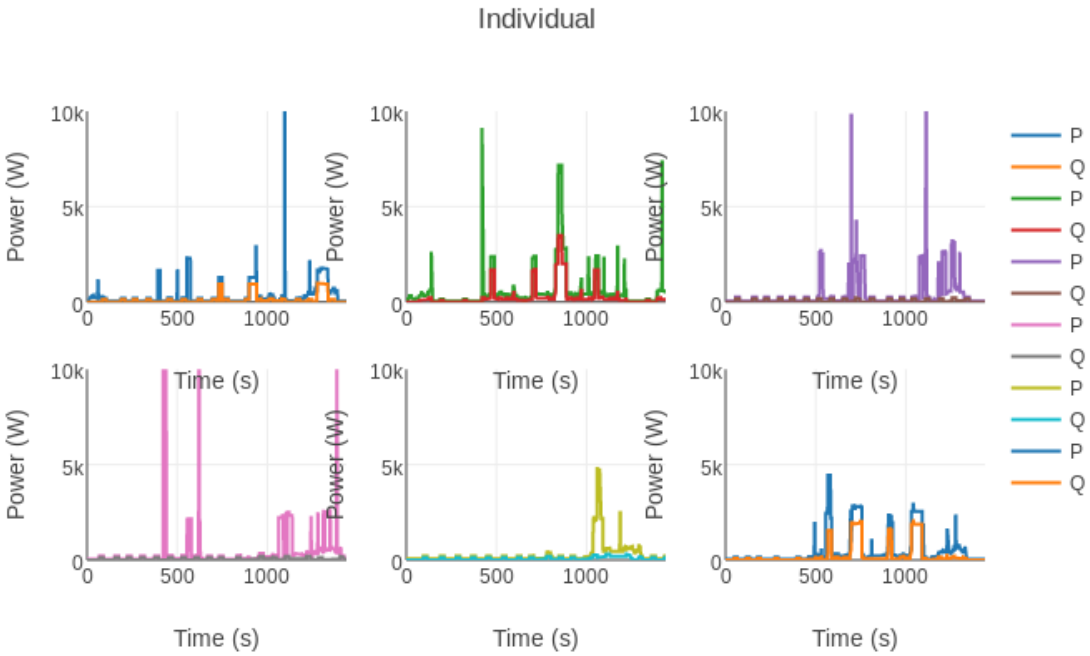


Fig. A.7 Individual loads of 6 profiles of bus 2 (with numbers 39, 42, 44 on the first row and 55, 60, 74 on the second row, left to right)

Appendix B

IEEE33 Bus Medium Voltage Network

The IEEE 33-bus medium voltage network is described in [5] and has a rated voltage of 12.66 kV. Branches and their resistance and reactance can be seen in the tables below.

From (bus nr)	To (bus nr)	R (pu)	X (pu)
1	2	0.0922	0.0470
2	3	0.4930	0.2511
3	4	0.3660	0.1864
4	5	0.3811	0.1941
5	6	0.8190	0.7070
6	7	0.1872	0.6188
7	8	0.7114	0.2351
8	9	1.0300	0.7400
9	10	1.0440	0.7400
10	11	0.1966	0.0650
11	12	0.3744	0.1238
12	13	1.4680	1.1550
13	14	0.5416	0.7129
14	15	0.5910	0.5260
15	16	0.7463	0.5450
16	17	1.2890	1.7210
17	18	0.7320	0.5740
2	19	0.1640	0.1565
19	20	1.5042	1.3554

Table B.1 Line data

From (bus nr)	To (bus nr)	R (pu)	X (pu)
20	21	0.4095	0.4784
21	22	0.7089	0.9373
3	23	0.4512	0.3083
23	24	0.8980	0.7091
24	25	0.8980	0.7011
6	26	0.2030	0.1034
26	27	0.2842	0.1447
27	28	1.0590	0.9337
28	29	0.8042	0.7006
29	30	0.5075	0.2585
30	31	0.9744	0.9630
31	32	0.3105	0.3619
32	33	0.3410	0.5302
8	21	2.0000	2.0000
9	15	2.0000	2.0000
12	22	2.0000	2.0000
18	33	0.5000	0.5000
25	29	0.5000	0.5000

Table B.2 Line data continued

Appendix C

Data

In this appendix, the mean and standard deviation of MAPE scores of each experiment are shown in a table.

C.1 State Estimation and Imputation Technique

Table C.1 Scores of varying state estimators and imputation techniques.

I	SE	Base				(Partly) Trusted				Untrusted			
		V		θ		V		θ		V		θ	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
None	IRWLS	0.00	0.00	0.00	0.00	-	-	-	-	-	-	-	-
	EKF	0.00	0.00	0.00	0.00	-	-	-	-	-	-	-	-
	UKF	0.00	0.00	0.00	0.00	-	-	-	-	-	-	-	-
LOCF	IRWLS	-	-	-	-	0.16	0.01	30.24	1.62	NC	NC	NC	NC
	EKF	-	-	-	-	0.09	0.02	17.86	2.67	0.48	0.08	92.15	7.26
	UKF	-	-	-	-	0.07	0.01	11.85	1.80	0.43	0.04	93.51	6.14
MEAN	IRWLS	-	-	-	-	0.23	0.01	43.54	1.13	NC	NC	NC	NC
	EKF	-	-	-	-	0.09	0.00	15.45	0.56	0.28	0.02	53.67	5.36
	UKF	-	-	-	-	0.07	0.00	11.51	0.45	0.29	0.02	53.37	4.31

I=Imputation Technique, SE=State Estimator, IRWLS=Iterative Recursive Weighted Least Squares,
EKF=Extended Kalman Filter, UKF=Unscented Kalman Filter,
LOCF=Last-observation-carried-forward, NC=No Convergence

C.2 Differential Privacy

Table C.2 Scores of varying the threshold.

t	(Partly) Trusted				Untrusted			
	V		θ		V		θ	
	μ	σ	μ	σ	μ	σ	μ	σ
0.01	0.044	0.001	9.04	0.20	4.57	0.31	259.96	21.00
0.05	0.046	0.001	9.31	0.25	6.25	0.42	468.60	32.70
0.1	0.048	0.001	9.70	0.22	5.13	0.28	434.30	26.80
0.2	0.052	0.001	10.77	0.26	2.38	0.19	240.35	13.43
0.3	0.056	0.002	11.60	0.33	1.38	0.14	164.33	12.06
0.4	0.059	0.002	12.28	0.44	0.90	0.08	121.32	8.40
0.5	0.060	0.002	12.60	0.46	0.63	0.06	94.18	7.85
1.0	0.064	0.002	13.45	0.44	0.28	0.02	51.06	3.09

Table C.3 Scores of varying the privacy budget of the conservative group.

ε_C	(Partly) Trusted				Untrusted			
	V		θ		V		θ	
	μ	σ	μ	σ	μ	σ	μ	σ
0.01	0.064	0.002	13.28	0.53	0.281	0.023	51.00	3.38
0.05	0.064	0.002	13.44	0.52	0.266	0.017	53.45	5.33
0.1	0.064	0.002	13.24	0.70	0.268	0.017	49.20	2.32
0.2	0.063	0.002	12.86	0.49	0.264	0.015	48.47	2.21
0.3	0.061	0.002	12.47	0.40	0.258	0.019	47.87	3.22
0.4	0.060	0.002	12.25	0.35	0.260	0.018	46.99	3.78
0.5	0.058	0.002	11.96	0.31	0.245	0.016	46.15	2.93

Table C.4 Scores of varying the privacy budget of the moderate group.

ε_M	(Partly) Trusted				Untrusted			
	V		θ		V		θ	
	μ	σ	μ	σ	μ	σ	μ	σ
0.05	0.065	0.003	13.72	0.50	0.266	0.017	53.31	3.61
0.1	0.065	0.002	13.44	0.41	0.268	0.017	52.86	3.42
0.15	0.064	0.002	13.35	0.62	0.264	0.015	51.57	2.70
0.20	0.064	0.002	13.47	0.44	0.258	0.019	50.81	2.77
0.25	0.063	0.002	13.25	0.50	0.260	0.018	50.37	3.96
0.30	0.063	0.002	13.13	0.44	0.245	0.016	49.27	2.89
0.35	0.063	0.002	12.96	0.50	0.245	0.016	49.14	3.52
0.40	0.061	0.002	12.76	0.44	0.245	0.016	48.76	3.54
0.45	0.061	0.002	12.76	0.40	0.245	0.016	47.99	3.06
0.50	0.061	0.002	12.66	0.35	0.245	0.016	48.53	3.83

Table C.5 Scores of varying the fraction of users in the conservative group.

ε_C	(Partly) Trusted				Untrusted			
	V		θ		V		θ	
	μ	σ	μ	σ	μ	σ	μ	σ
0.10	0.051	0.001	10.43	0.29	0.225	0.013	44.62	3.13
0.20	0.052	0.001	10.86	0.29	0.248	0.017	49.55	3.92
0.30	0.054	0.001	11.25	0.27	0.274	0.020	54.58	4.68
0.40	0.056	0.002	11.74	0.28	0.324	0.029	69.96	6.94
0.50	0.063	0.002	12.92	0.48	0.343	0.039	71.72	7.92
0.54	0.064	0.002	13.20	0.56	0.380	0.039	81.42	9.33
0.60	0.068	0.002	14.18	0.65	0.426	0.057	88.36	10.81

Table C.6 Scores of varying the composition size.

k	(Partly) Trusted				Untrusted			
	V		θ		V		θ	
	μ	σ	μ	σ	μ	σ	μ	σ
1	0.065	0.003	13.51	0.50	0.278	0.017	51.95	3.46
2	0.096	0.003	19.15	0.82	0.695	0.055	105.23	6.80
5	0.161	0.007	31.49	1.47	NC	NC	NC	NC
15	0.452	0.046	73.92	6.73	NC	NC	NC	NC
30	1.432	0.137	170.47	11.86	NC	NC	NC	NC
60	6.148	0.641	547.99	58.59	NC	NC	NC	NC

C.3 Context

Table C.7 Scores of varying the number of dwellings per substation.

N	(Partly) Trusted				Untrusted			
	V		θ		V		θ	
	μ	σ	μ	σ	μ	σ	μ	σ
10	0.248	0.020	109.58	10.87	0.603	0.076	241.48	24.65
25	0.154	0.007	39.80	2.21	0.601	0.069	139.18	15.94
50	0.096	0.003	22.74	0.81	0.378	0.036	78.90	5.22
75	0.076	0.003	16.72	0.7	0.326	0.029	61.53	3.51
100	0.064	0.002	13.28	0.44	0.280	0.025	51.28	3.51

Table C.8 Scores of varying the month and day of the week.

d	m	(Partly) Trusted				Untrusted			
		V		θ		V		θ	
		μ	σ	μ	σ	μ	σ	μ	σ
Weekday	January	0.065	0.004	11.10	0.65	0.213	0.016	28.43	1.75
	February	0.065	0.005	12.94	0.83	0.218	0.017	34.38	2.67
	March	0.063	0.005	11.80	0.66	0.208	0.019	31.90	2.35
	April	0.066	0.006	14.35	1.42	0.303	0.016	54.67	2.65
	May	0.067	0.006	14.65	1.05	0.270	0.022	51.18	3.39
	June	0.076	0.005	31.38	2.61	0.335	0.019	116.18	6.61
	July	0.076	0.006	31.43	2.79	0.338	0.023	116.11	9.09
	August	0.064	0.005	29.00	2.21	0.283	0.018	106.98	8.37
	September	0.070	0.002	14.97	0.60	0.304	0.025	56.19	3.44
	October	0.070	0.002	13.33	0.48	0.310	0.019	53.97	3.69
	November	0.070	0.002	12.15	0.56	0.274	0.014	45.84	43.90
	December	0.073	0.002	11.67	0.53	0.316	0.023	45.48	2.17
Weekend	January	0.065	0.004	10.31	0.46	0.213	0.016	39.34	2.51
	February	0.065	0.005	10.53	0.34	0.218	0.017	40.75	2.55
	March	0.063	3.005	10.53	0.39	0.208	0.019	41.55	1.65
	April	0.065	0.006	11.89	0.40	0.303	0.016	46.67	2.44
	May	0.067	0.006	14.56	0.65	0.277	0.022	54.13	2.41
	June	0.076	0.006	27.36	0.65	0.335	0.019	111.77	8.18
	July	0.076	0.006	26.33	0.89	0.338	0.023	103.85	9.33
	August	0.064	0.005	27.60	0.72	0.283	0.018	113.06	8.38
	September	0.070	0.002	15.86	0.61	0.304	0.025	64.65	3.56
	October	0.070	0.002	10.69	0.50	0.310	0.020	40.33	2.06
	November	0.067	0.002	10.77	0.35	0.274	0.014	41.91	2.94
	December	0.073	0.002	10.01	0.31	0.316	0.023	38.00	2.27

d=Day of the Week, m=month