# ANOMALY DETECTION FRAMEWORKS FOR IDENTIFYING ENERGY THEFT AND METER IRREGULARITIES IN SMART GRIDS

YIP SOOK CHIN

INSTITUTE FOR ADVANCED STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

University of Malaya

# ANOMALY DETECTION FRAMEWORKS FOR IDENTIFYING ENERGY THEFT AND METER IRREGULARITIES IN SMART GRIDS

**YIP SOOK CHIN**

**THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**INSTITUTE FOR ADVANCED STUDIES**
**UNIVERSITY OF MALAYA**
**KUALA LUMPUR**

**2019**

University of Malaya

**UNIVERSITI MALAYA**

**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: **YIP SOOK CHIN**

Registration/Matric No.: **HHD140002**

Name of Degree: **Doctor of Philosophy (Ph.D)**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

**ANOMALY DETECTION FRAMEWORKS FOR IDENTIFYING ENERGY THEFT AND METER IRREGULARITIES IN SMART GRIDS**

Field of Study: **POWER SYSTEM PROTECTION**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

University of Malaya

**ANOMALY DETECTION FRAMEWORKS FOR IDENTIFYING ENERGY THEFT AND METER IRREGULARITIES IN SMART GRIDS**

## ABSTRACT

Non-technical losses including electricity theft and anomalies in meter readings are estimated to cost the utility providers losses of approximately $96 billion per annum globally. Although the implementation of smart grids offers technical and social advantages, the smart meters deployed in advanced metering infrastructure are susceptible to more sophisticated types of malicious attack as compared to conventional mechanical meters. To curb non-technical losses, utility providers are increasingly leveraging on real-time smart metering to identify theft and meter irregularities. In the first part of this study, a linear regression-based anomaly detection framework is put forward to study consumers' energy utilization behavior and evaluate their anomaly coefficients to detect the localities of the compromised and defective smart meters. The main idea is to model the amount of stolen energy at a smart meter as an anomaly coefficient. Specifically, a zero-anomaly coefficient indicates a faithful meter while a non-zero one represents an anomalous/defective meter. However, some of the predicted elements of anomaly coefficient vector might show inaccurate values when energy theft/meter irregularities take place only during a certain period in a day. Thus, categorical variable and detection coefficient are introduced in the framework to identify the periods and localities of consumers' malfeasance/faulty meters. By investigating the anomaly coefficients and detection coefficients, non-technical losses can be deduced whether they occur either all the time or only during a certain period in a day. However, the linear regression-based framework assume that power line losses are known. Therefore, in the second part of this study, the assumption of known power line losses is relaxed, and a new anomaly detection framework is designed to take into consideration

University of Malaya

the impact caused by technical losses and measurement noise. Similarly, the goal is to identify anomalous consumption patterns within the billing reports transmitted to utility provider by evaluating consumers' anomaly coefficients. To improve detection accuracy and reduce false positives, metrics known as loss factor and error term are introduced. Linear programming is utilized to solve for anomaly coefficients and loss factors by minimizing the error terms. The linear programming-based anomaly detection framework can still detect irregularities in meter readings regardless of whether non-technical losses occur all the time or at varying rates during intermittent intervals in a day. In addition, it can estimate the percentage of technical losses based on measurements at the data collector and the knowledge of the distribution network. To evaluate the performance of the proposed frameworks, a diverse set of non-technical loss attack functions is investigated and generated such that the experiments are closely related to the possible real-world energy fraud/meter irregularities scenarios. Subsequently, an advanced metering infrastructure test rig is constructed in the laboratory to validate the reliability and performance of both anomaly detection frameworks. Results from simulations and test rig show that both anomaly detection frameworks can reveal the amount of under-reporting/over-reporting by smart meters based on a small volume of consumers' energy consumption data samples regardless of the type of consumer, thereby reducing loss incurred due to non-technical losses.

**Keywords:** anomaly detection, non-technical losses, AMI, linear regression, linear programming.

University of Malaya

**RANGKA KERJA PENGESANAN ANOMALI UNTUK MENGENALPASTI KECURIAN TENAGA ELEKTRIK DAN PENYELEWENGAN METER DALAM GRID PINTAR**

**ABSTRAK**

Kehilangan tenaga bukan teknikal termasuk kecurian tenaga dan penyelewengan meter dianggarkan telah menyebabkan pembekal utiliti di seluruh dunia mengalami kerugian sebanyak $96 bilion setiap tahun. Walaupun pelaksanaan grid pintar menawarkan kelebihan dari segi teknikal dan sosial, meter pintar yang digunakan dalam pemeteran infrastruktur maju mudah terdedah kepada serangan dan pencerobohan rangkaian yang lebih canggih berbanding dengan meter mekanikal konvensional. Untuk membendung kehilangan tenaga bukan teknikal, pembekal utiliti menggunakan pemeteran pintar masa nyata untuk mengenalpasti lokasi kecurian tenaga dan penyelewengan meter. Dalam fasa pertama penyelidikan ini, rangka kerja pengesanan anomali berasaskan regresi linear dikemukakan untuk mengkaji corak penggunaan tenaga pengguna dan menilai pekali anomali mereka untuk mengenalpasti meter pintar yang dikompromi/rosak. Idea utama adalah untuk memodelkan jumlah tenaga yang dicuri dari meter pintar sebagai pekali anomali. Khususnya, pekali anomali sifar menunjukkan meter yang normal manakala pekali anomali yang bukan sifar mewakili meter yang dikompromi/rosak. Walau bagaimanapun, beberapa elemen dalam vektor pekali anomali mungkin akan menunjukkan nilai yang tidak tepat apabila kecurian tenaga/kerosakan meter hanya berlaku dalam tempoh tertentu dalam sehari. Oleh itu, pembolehubah mutlak dan pekali pengesanan diperkenalkan dalam rangka kerja tersebut untuk mengenalpasti tempoh dan lokasi kecurian tenaga/kerosakan meter. Dengan merujuk kepada pekali anomali dan pekali pengesanan, kesimpulan bahawa kehilangan tenaga bukan teknikal berlaku sama ada sepanjang masa atau hanya

University of Malaya

dalam tempoh tertentu dalam sehari boleh dibuat. Dalam penyelidikan ini, rangka kerja yang berasaskan regresi linear mengandaikan bahawa nilai kehilangan tenaga secara teknikal telah diketahui. Dalam fasa kedua penyelidikan ini, andaian tersebut dilonggarkan dan sebuah rangka pengesanan anomali baharu dikemukakan untuk mengambil kira kesan kehilangan tenaga secara teknikal dan ralat pengukuran dalam analisis pengesanan. Seperti penyelidikan fasa pertama, matlamat penyelidikan adalah untuk mengenalpasti corak anomali penggunaan tenaga dengan menilai pekali anomali pengguna. Untuk meningkatkan ketepatan pengesanan dan mengurangkan positif palsu, faktor kehilangan dan istilah ralat diperkenalkan. Pengaturcaraan linear digunakan untuk menyelesaikan masalah pekali anomali dan faktor kehilangan dengan meminimumkan istilah ralat. Rangka kerja pengesanan anomali berdasarkan pengaturcaraan linear masih dapat mengesan penyelewengan meter tanpa mengira kehilangan tenaga bukan teknikal berlaku sama ada sepanjang masa atau pada kadar yang berbeza dalam selang intermiten sepanjang sehari. Rangka kerja pengesanan anomali itu juga dapat menganggarkan peratusan kehilangan tenaga secara teknikal berdasarkan pengukuran di pengumpul data dan pengetahuan tentang rangkaian pengedaran tenaga. Untuk menilai prestasi rangka kerja, pelbagai jenis fungsi serangan kehilangan tenaga bukan teknikal telah dikaji agar eksperimen yang dijalani berkait rapat dengan scenario kecurian tenaga/kerosakan meter yang sebenar. Sebuah rig ujian pemeteran infrastruktur maju telah dibina di dalam makmal untuk mengesahkan kebolehpercayaan dan prestasi rangka kerja pengesanan anomali. Keputusan dari simulasi dan rig ujian menunjukkan bahawa rangka kerja tersebut dapat mendedahkan jumlah tenaga yang tidak/terlebih dilaporkan berdasarkan sampel data yang kecil, tanpa mengira jenis pengguna dan dapat mengurangkan kerugian akibat kehilangan tenaga bukan teknikal.

**Kata kunci:** pengesanan anomali, kehilangan tenaga bukan teknikal, AMI, regresi linear, pengaturcaraan linear.

University of Malaya

## ACKNOWLEDGEMENTS

First of all, I would like to extend my most sincere gratitude to my supervisors Dr. Tan Chia Kwang, Assoc. Prof. Dr. Wong Kok Sheik (Monash University Malaysia) and Prof. Dr. Raphael Phan Chung Wei (Multimedia University Cyberjaya) and the late Prof. Dr. Hew Wooi Ping for their invaluable guidance and assistance throughout the course of my Ph.D. studies. They have always been encouraging, helpful and giving good advice whenever needed. This study would never have been completed without their support and dedicated involvement in every research stage.

I would like to express my greatest appreciation to my fellow colleagues at the Faculty of Engineering of MMU for their support in this research. In particular, I am deeply grateful to my sifu, Mr. Gan Ming Tao. I wish to thank him for his patience in explaining the research ideas in detail and guiding me in developing the algorithms from scratch. My sincere thanks also goes to my guru, Dr. Tan Wooi Nee for her dedicated guidance and support. She has offered insightful comments, suggestions and discussions that undoubtedly helped to facilitate my research progress.

Besides, I would like to express my gratitude to Assoc. Prof. Dr. Lee Ching Kwang for his unfailing support and assistance. I am also indebted to my good friends in MMU, including Dr. Vishnu Monn Baskaran, Dr. Ivan Ku, Dr. Choo Kan Yeep, Dr. Guo Xiao Ning, Mr. Lo Yew Chiong and Dr. Goh Vik Tor for their company and encouragement throughout this journey.

Last but not the least, I owe my deepest gratitude to my family members, specially my ever supporting husband, Alfred Yap, my supportive siblings, my loving parents and in laws for all of the sacrifices that they have made on my behalf.

University of Malaya

## TABLE OF CONTENTS



University of Malaya

University of Malaya

University of Malaya

University of Malaya

University of Malaya

## LIST OF FIGURES



University of Malaya

University of Malaya

University of Malaya

University of Malaya

University of Malaya

## LIST OF TABLES



University of Malaya

University of Malaya

## LIST OF SYMBOLS AND ABBREVIATIONS

| | | |
|---|---|---|
| $E_{t_i}^d$ | : | error term at time slot $t_i$ on day $d$. |
| $\beta_n$ | : | detection coefficient of consumer $n$. |
| $\gamma$ | : | faulty SMs. |
| $\lambda$ | : | TLs. |
| $\theta$ | : | energy theft. |
| $a_n$ | : | anomaly coefficient of consumer $n$. |
| $a_{t_i,n}$ | : | anomaly coefficient of consumer $n$ at time interval $t_i$. |
| $c_{t_i}$ | : | measurement of data collector at time interval $t_i$. |
| $g_{t_i}$ | : | total energy generated by all the consumers at time interval $t_i$. |
| $l_{t_i}^d$ | : | loss factor at time interval $t_i$ on day $d$. |
| $t_i$ | : | time interval. |
| $x_n$ | : | categorical variable of consumer $n$. |
| $y_{t_i}$ | : | meter discrepancy at time interval $t_i$. |
| $y_{t_i}^d$ | : | meter discrepancy at time interval $t_i$ on day $d$. |
| ADF | : | Anomaly Detection Framework. |
| AI | : | Artificial Intelligence. |
| AMI | : | Advanced Metering Infrastructure. |
| AMIDS | : | AMI Intrusion Detection System. |
| ARIMA | : | Auto-Regressive Integrated Moving Average. |
| ARMA | : | Auto-Regressive Moving Average. |
| CPU | : | central processing unit. |
| CT | : | current transformer. |
| CVLR-ETDM | : | Categorical Variable-enhanced LR-ETDM. |
| DER | : | distributed energy resource. |
| DG | : | distributed generation. |
| DR | : | detection rate. |
| DS | : | distribution substation. |
| DSM | : | demand side management. |
| DT | : | distribution transformer. |
| ELM | : | Extreme Learning Machine. |
| ESS | : | energy storage system. |
| FP | : | false positives. |
| GPRS | : | General Packet Radio Service. |
| GSM | : | Global System for Mobile communications. |
| HAN | : | Home Area Network. |
| HV | : | high voltage. |
| ICT | : | information and communication technology. |

University of Malaya

| | | |
|---|---|---|
| IoT | : | Internet of Things. |
| KLD | : | Kullback-Leibler Divergence. |
| kWh | : | kilowatt-hour. |
| LP | : | Linear Programming. |
| LR | : | Linear Regression. |
| LR-ETDM | : | Linear Regression-based scheme for Detection of Energy Theft and Defective Smart Meters. |
| LSE | : | linear system of equations. |
| LUD | : | Lower-Upper Decomposition. |
| LV | : | low voltage. |
| MCBs | : | Miniature Circuit Breakers. |
| MDMS | : | Meter Data Management Systems. |
| MLR | : | Multiple Linear Regression. |
| NAN | : | Neighborhood Area Network. |
| NILM | : | non-intrusive load monitoring. |
| NTL | : | Non-Technical Loss. |
| NTLs | : | Non-Technical Losses. |
| PEA | : | Provincial Electricity Authority. |
| PHEV | : | plug-in hybrid electric vehicle. |
| PLC | : | Power Line Carrier. |
| PMUs | : | phasor measurement units. |
| PVC | : | Polyvinyl chloride. |
| RF | : | radio frequency. |
| RFID | : | radio-frequency identification. |
| SCADA | : | Supervisory Control and Data Acquisition. |
| SG | : | Smart Grid. |
| SGs | : | Smart Grids. |
| SM | : | Smart Meter. |
| SME | : | Small and Medium Enterprises. |
| SMs | : | Smart Meters. |
| SVM | : | Support Vector Machine. |
| TCP/IP | : | Transmission Control Protocol/Internet Protocol. |
| TL | : | Technical Loss. |
| TLs | : | Technical Losses. |
| TNB | : | Tenaga Nasional Berhad. |
| TOU | : | time-of-use. |
| UP | : | utility provider. |
| UPs | : | utility providers. |
| W | : | watts. |
| WAN | : | Wide Area Network. |

University of Malaya

## LIST OF APPENDICES



University of Malaya

# CHAPTER 1: INTRODUCTION

In this chapter, an overview of the thesis is presented. It includes the research problem statements, objectives, scopes, limitations and contributions, under the general topic of anomaly detection. Then, the structure of the thesis is briefly delineated to aid the understanding of the presentation.

## 1.1 Motivation of Anomaly Detection

Non-technical losses (NTLs), which includes meter irregularities, non-payment by consumers and errors in record-keeping, is a daunting problem in electric power systems since the early days of energy billings. Interestingly, the main contributor of NTLs is energy theft which causes severe impacts for both utility providers (UPs) and legitimate consumers, resulting in a total of staggering $96 billion lost every year globally (Northeast Group, 2017). Specifically, the latest estimate shows that UPs in the United States alone lost billions of dollars in revenue annually (McDaniel & McLaughlin, 2009), while energy theft in developing countries amounts to approximately half of the total energy delivered (Pedro, 2009). NTLs not only result in excessive energy usage which may cause detrimental electrical system failures or power surges (Foster, 2017), they also indirectly encourage fraudulent activities such as unauthorized growing of controlled drugs (Accenture, 2011). UPs always amortize NTLs by rising energy charges on lawful consumers. Specifically, the consumers being billed for legal consumption and regularly paying their bills are unknowingly subsidizing the energy thieves who do not pay for electricity consumption. As mentioned by Smith (2004); Pedro (2009); Refou, Alsafasfeh, and Alsoud (2015), the amount of energy loss in the distribution grids varies between 7 and 50% of the total supplied energy (subject to the characteristics of the distribution network and country), which undeniably justifies the strong efforts that UPs and government are investing towards

University of Malaya

identifying and inspecting anomalous consumption trends to ultimately avoid significant economic losses.

Today, the deployment of advanced metering infrastructure (AMI) in Smart Grids (SGs) has become significantly crucial for mitigating NTLs and providing better power quality, higher reliability, more accurate billing as well as lower utility costs. In spite of these societal and technical advantages, AMI is still susceptible to more sophisticated types of malicious attack. Specifically, Smart Meters (SMs) endowed in AMI enable features such as remote update of firmware and automatic transmission of metering data. Nevertheless, these functionalities indirectly create a "back door" for malicious consumers. For example, an energy thief can easily obtain the root access of the SMs to manipulate the energy consumption readings (McLaughlin, Podkuiko, & McDaniel, 2010). Subsequently, the misinformation of energy usage might severely damage the electrical power infrastructure when attacks are injected into the control systems. Considerable energy fraud and incorrect usage information might also delude the UPs into making wrong decisions about domestic/regional capacity and consumption. These inaccurate information might conceal the upcoming problems or ongoing attacks from UPs. In such a case, criminals or terrorists can easily misuse such facilities to launch massive detrimental attacks on local or national infrastructure. All these security challenges definitely hamper consumers' trust in adopting SGs as privacy and safety of their data are not guaranteed (Engel, 2013).

Therefore, implementing anomaly detection frameworks to detect energy theft and meter irregularities in SGs is imperative to address NTLs. It is also important in establishing consumers' trust in adopting SGs to replace its antique predecessor.

University of Malaya

## 1.2 Problem Statements

To curb NTLs, UPs are gradually leveraging on data analytics and real-time smart metering in AMI to detect localities with high probability of energy theft/meter irregularities. In general, several issues have been brought to the researchers' attention:

1. Most existing classification-based NTL detection schemes require long-term measurement and monitoring before anomaly detection can be executed precisely. The large sample size requirement naturally results in longer detection delay.
2. Some of the existing detection schemes are susceptible to contamination attacks. In other words, an energy thief can simply deceive the learning machine to accept an anomalous pattern as a normal through granular changes in data and data pollution.
3. Non-malicious factors can change the energy usage trend and hence might affect the performance of some classification-based detection analysis.
4. Some of the detection schemes are highly dependent on the historical dataset. Lack of a thorough dataset of attack samples limits the detection rate.

These issues might result in high false positive rate if they are not properly dealt with. This in turn increases the overall operation costs of UPs.

## 1.3 Research Objectives

This thesis aims to address the aforementioned problem statements and reduce losses incurred due to non-technical loss (NTL) activities.

The main objectives of this research are set out as follows:

1. To investigate and generate a diverse set of NTL attack functions such that it closely relates to the possible real-world AMI energy thefts/meter irregularities scenarios.

University of Malaya

2. To design anomaly detection frameworks that can effectively detect the localities of energy theft and meter irregularities in smart grids using linear regression and linear programming techniques.
3. To evaluate the proposed anomaly detection frameworks using smart energy data from the Irish Smart Energy Trial.
4. To design and build an AMI test rig in the laboratory to further validate the reliability and performance of the entire anomaly detection framework in real smart grid environment.

## 1.4 Research Scopes and Limitations

There are several scopes and restrictions that needed to be highlighted throughout the research duration, in order to conduct research efficiently and achieve research objectives. The scopes and restrictions are prescribed as follows:

1. Since energy theft samples in SGs are rare, the assessment of the proposed anomaly detection frameworks is performed by simulating NTL attacks, whereby consumers' benign readings in the smart energy dataset are modified.
2. This thesis focuses on the NTL detection in low voltage (LV) distribution network, which include: residential, commercial and light industrial consumers by using the half-hourly smart energy data from the Irish Smart Energy Trial as well as data from the test rig.
3. The logistics of how an energy thief can modify the communication signals is not a focus of this thesis. The goal is to identify malicious attacks under the assumption that the energy thief has successfully compromised the integrity of Smart Meter (SM) readings.
4. It is assumed that the data collector can be trusted (i.e., not tampered). This

University of Malaya

assumption is easily justified when the data collector is placed in a distribution substation (DS) on the same premise as the operation center.

5. It is assumed that all consumers' premises are equipped with a SM. Therefore, the impact caused by consumers without a SM is not considered.

## 1.5 Research Contributions

The main contributions of this research are summarized as follows.

1. Advances the research in anomaly detection of detecting the localities of under-reporting and over-reporting by SMs based on linear regression and linear programming techniques for efficient detection and classification of NTL activities.
2. Reduces false positives and improves detection accuracy by taking into consideration the impact caused by measurement noise and technical losses (TLs) on the detection framework for more cost-effective anomaly detection.
3. Realizes greater flexibility, faster and enhanced practicality in the detection of energy theft/defective meters based on a small volume of consumers' energy consumption data samples regardless of the types of consumer and the amount of technical losses (TLs). The proposed framework can be scaled to accommodate anomaly detection for more consumers.

## 1.6 Structure of Thesis

Generally, the structure of this thesis is divided into two categories, namely: (i) project methodology and; (ii) research methodology. The project methodology consists of the overall work completed for developing the anomaly detection frameworks. The following subsections give a brief description of project and research methodologies of this thesis.

University of Malaya

### 1.6.1 Project Methodology

This thesis is proposed in an effort to minimize the losses and costs incurred due to NTL events in the LV distribution network, which are estimated to be approximately 20% throughout Peninsular Malaysia (Nagi, Yap, Tiong, Ahmed, & Mohamad, 2010). The overall project methodology is shown in Figure 1.1.

The development of the anomaly detection frameworks, namely the Linear Regression (LR)-based and Linear Programming (LP)-based anomaly detection frameworks, are the main focuses of this research study. As illustrated in Figure 1.1, the procedures involved in the development of the anomaly detection frameworks includes: Conduct extensive literature review to investigate various types of NTL attack function used to defraud the UPs including those schemes that are widely deployed in both SGs and conventional power grids; investigate state-of-the-art schemes proposed for mitigating NTL events and identifying malicious consumers; study the advantages, technologies and challenges involved in the design and deployment of SGs; carry out data collection and load profile inspection for detection of normal and malicious energy consumption behavior; design *anomaly coefficient* and *detection coefficient* for *classification* of consumers (i.e., either honest or anomalous); look into the impact caused by TLs and measurement noise/error on the detection analysis; finally perform frameworks testing and validation to confirm the reliability of the proposals.

### 1.6.2 Research Methodology

The research methodology is put forward in order to develop two intelligent anomaly detection frameworks for detection of NTL events such as energy theft and meter irregularities in SGs, as presented in Figure 1.2. The research methodology is embedded within the project methodology shown in Figure 1.1.

In this thesis, the smart energy data from the Irish Smart Energy Trial (Commission for

University of Malaya

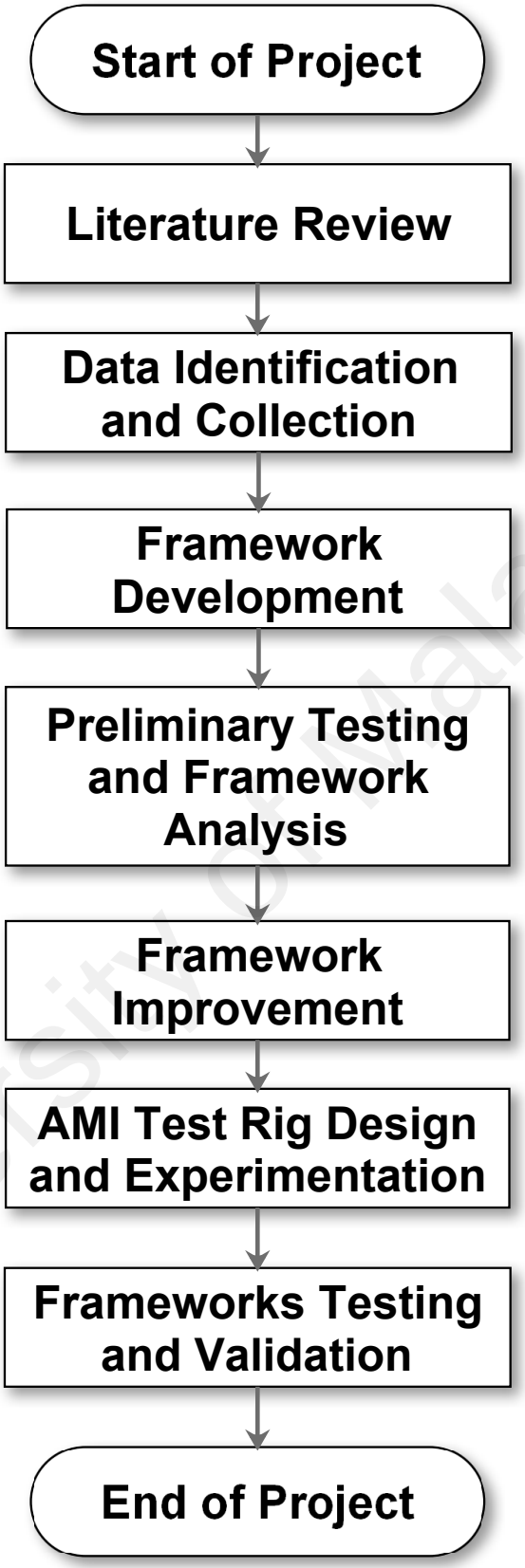



**Figure 1.1: Flow chart of the proposed project methodology for detection of NTL events.**

University of Malaya

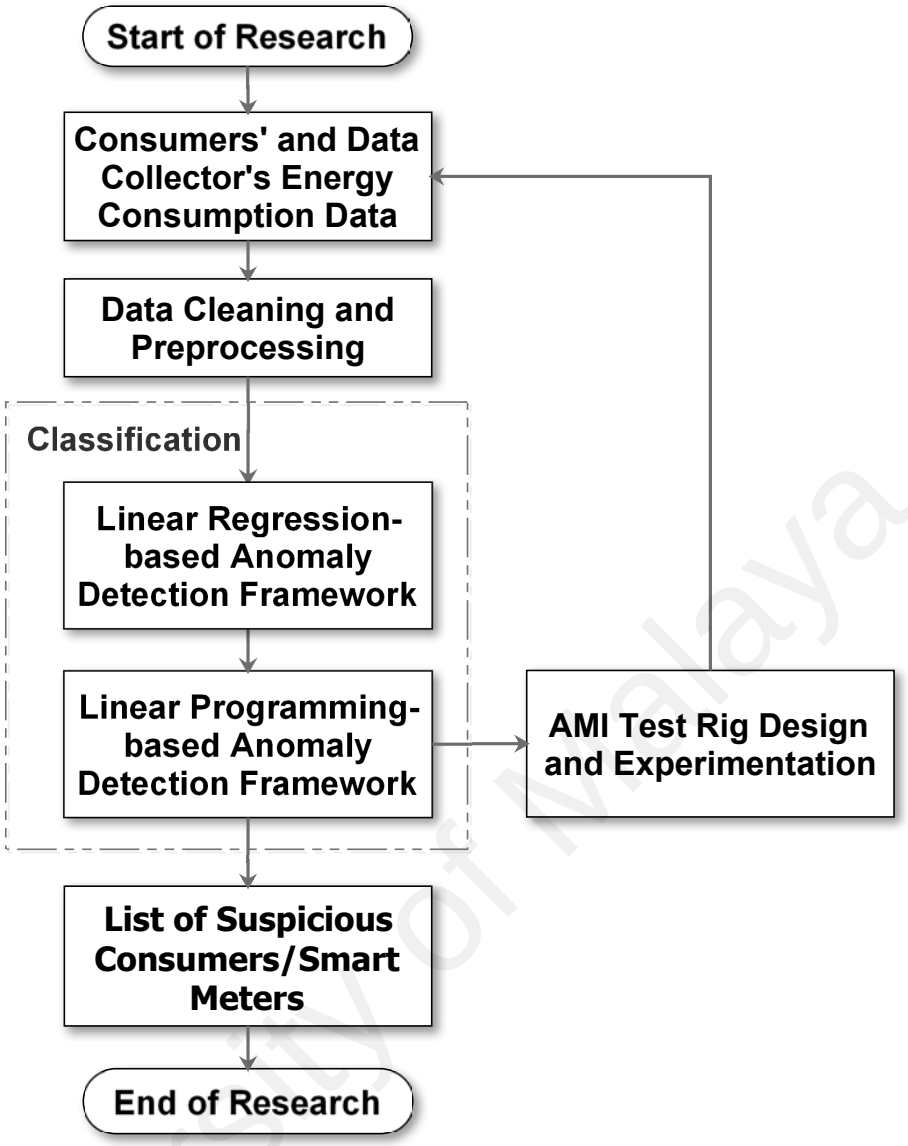



**Figure 1.2: Flow chart of the proposed research methodology for detection of NTL events.**

Energy Regulation, 2009) is first used to study consumers' energy consumption trend for revealing the localities of potential energy frauds and faulty meters. The dataset consists of 30-minute energy consumption reports for both Irish commercial and residential premises of different contracted power between 2009 and 2010. The data are transformed into the required format for Multiple Linear Regression (MLR), by performing data cleaning and preprocessing. Then, the detection and classification are undertaken by the proposed LR-based anomaly detection framework, which is the intelligent "anomaly detection engine". The LR-based anomaly detection framework detect NTL events based on the

University of Malaya

energy balance analysis. Particularly, the proposed frameworks shortlist locations with high probability of malicious activities according to the meter discrepancies at the DS and model the amount of under-reporting/over-reporting by a SM as an *anomaly coefficient*. By solving the anomaly coefficient of each consumer, the localities of the suspicious energy thieves and malicious SMs can be detected.

Next, in the pursuit of higher detection rate and lower false positives, a LP-based anomaly detection framework is designed. Apart from detecting the localities of fraudulent consumers and malicious SMs, the LP-based anomaly detection framework achieves significant improvement in detection rate and false positive reduction by looking into the impact caused by measurement noise/error and TLs on the detection analysis. Moreover, the proposed framework is also able to estimate the percentage of TLs based on measurements at the data collector and the knowledge of the distribution network.

Nevertheless, the DS SM readings as well as real energy theft sample do not exist in Malaysia because SGs are not fully deployed. Thus, an AMI test rig is designed and built in the laboratory to evaluate the reliability and performance of the proposed anomaly detection frameworks for identifying energy fraud and meter irregularities in real Smart Grid (SG) environment. Finally, the list of suspicious consumers/SMs is used by the UPs to plan for their NTL inspection activities.

### 1.6.3 Organization of the Thesis

The contents for each chapter, except for Chapter 1, are outlined briefly in this section. Chapter 2 provides the preliminary studies of SGs, literature review on different types of electricity loss, which include TLs and diverse types of NTL attack as well as existing NTL detection schemes. In Chapter 3, a novel linear regression-based anomaly detection framework is proposed to reveal energy theft and meter irregularities, regardless of whether NTLs take place only during a certain period in a day or all the time. Chapter 4 puts

University of Malaya

forward a linear programming-based anomaly detection framework, which takes into consideration the effect of measurement noise/error and TLs on detection analysis to reduce false positives and enhance detection accuracy. The proposed framework is then improved so that it can detect intermittent metering defects/energy fraud. To further validate the reliability and performance of the entire anomaly detection framework both in simulation and real SG environment, Chapter 5 describes the data collection and test setup. In Chapter 6, the performance of the proposed anomaly detection frameworks is assessed and discussed in both constant-rate and varying-rate cheating/malfunctioning scenarios. In addition, the strengths and weaknesses for each detection framework are investigated. Chapter 7 draws some conclusions and suggests possible future research directions.

University of Malaya

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Overview

The chapter serves to provide a review of important aspects in SG and AMI, particularly NTLs in SG environment. Firstly, fundamental studies on SGs and various aspects of AMI, i.e., subsystems, applications, challenges and security issues arising from AMI, are conducted in Section 2.2. In Section 2.3, consideration is given to the background and theoretical concepts pertaining to electricity losses that UPs experience, which include TLs and NTLs. Diverse types of NTL event, such as non-payment by consumers, meter irregularities and energy theft are also identified and presented in Section 2.3.2. Also considered are the diverse methods of energy theft that have commonly been exploited to steal energy from the electrical power distribution system. Next, some background issues concerning NTL detection schemes used in energy industry are reviewed in Section 2.4. Finally, the problems faced by the surveyed literature are identified and discussed.

## 2.2 Smart Grid Fundamentals

In recent years, the antiquated electrical power infrastructure that supplies energy to both residential and commercial premises is gradually being replaced with a set of digital systems known as the SGs. SGs are the next-generation power grids in which the energy management and distribution are upgraded by incorporating bi-directional information and communication technology (ICT) and pervasive computing capabilities for better control, efficiency, reliability and safety (Yan, Qian, Sharif, & Tipper, 2013). According to the United States Department of Energy's modern grid initiative (U.S. Department of Energy, 2008), SGs incorporate control methods, integrated communications and advanced sensing technologies into the current electrical power infrastructure. These modernized grids enhance consumers' and UPs' ability to monitor, control and forecast energy usage. As

University of Malaya

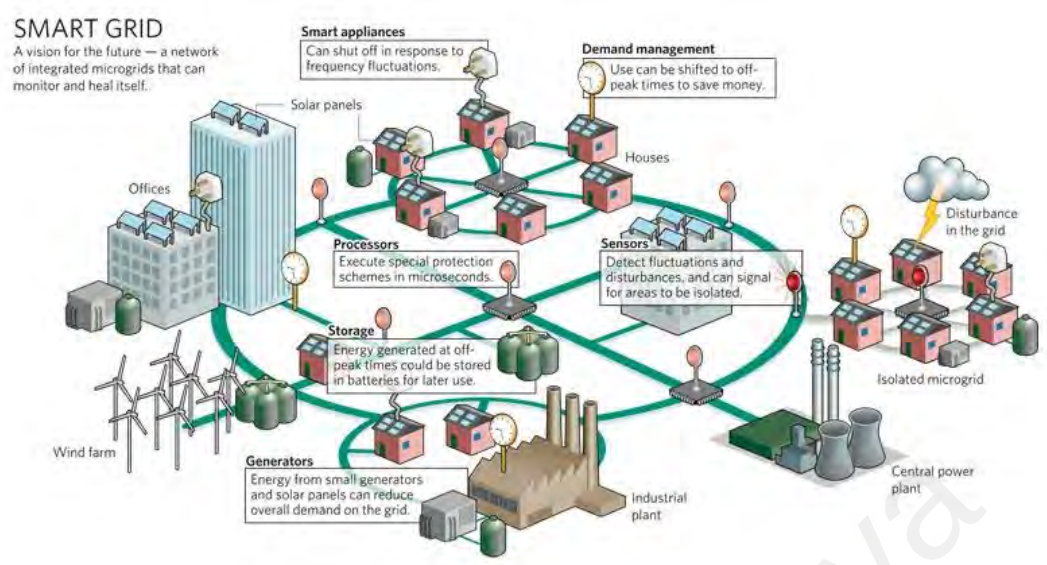



**Figure 2.1: The future power grid: smart grid (Marris, 2008).**

shown in Figure 2.1, SGs integrate microgrids, diverse distributed energy resource (DER) such as solar, wind and energy storage system (ESS) into a smart self-healing grid system to address existing energy management issues. In SGs, UPs encourage consumers to monitor and control their energy consumption by introducing demand side management (DSM) as incentive to promote a less fluctuating consumption pattern. In such a case, the consumers are financially motivated to shift their energy load to off-peak periods, and thereby reducing the fluctuation in the rate of energy consumption and peak-to-average ratio of the total energy demand. Therefore, the government, academia and energy industry are propelled to implement SGs to reduce greenhouse gas emissions, combat global warming and reach national energy independence (McDaniel & McLaughlin, 2009). Section 2.2.1 introduces the technologies of AMI, as the base of SGs, which is in charge of collecting information and data from consumers and loads.

### 2.2.1 Overview of AMI

Deploying an AMI is an essential step in the modernization of power grid. AMI offers an intelligent metering framework to fulfill one of the key initiatives of SGs– motivation

University of Malaya

and participation of the consumers (U.S. Department of Energy, 2008). AMI provides consumers with the information to make better consumption decisions and a variety of options to execute those decisions. On the other hand, AMI helps UPs improve consumer service by refining asset management processes and utility operations based on the smart metering data. AMI sends information of power-related events and consumers' energy consumption to both the consumers and UPs. Therefore, all parties can take part in the billing and peak demand reduction as well as make informed decisions in SGs (McLaughlin, Holbert, Fawaz, Berthier, & Zonouz, 2013). In short, AMI provides an important linkage between the UPs, consumers, generation and storage resources through the integration of numerous technologies such as integrated communications, smart metering, Home Area Network (HAN), software interfaces with existing utility operations and data management applications. In this thesis, only issues associated with utilization of AMI in electrical power distribution system are discussed.

### 2.2.2 Subsystems of AMI

AMI is an advanced framework which includes smart devices (i.e., SMs, data collectors and Internet of Things (IoT) devices), data collection platform (i.e., Meter Data Management Systems (MDMS) and head end) and different communication networks to integrate the collected data into software platforms and hardware interfaces (Rashed Mohassel, Fung, Mohammadi, & Raahemifar, 2014). In AMI, communication and electrical flows are bi-directional and overlay each other (Fang, Misra, Xue, & Yang, 2012). The architecture of electrical network and AMI in SG is presented in Figure 2.2.

#### 2.2.2.1 Smart Devices

Smart devices comprise of state-of-the-art software and hardware which are capable of collecting and measuring data at preset time stamp. These digitized devices are configured

University of Malaya

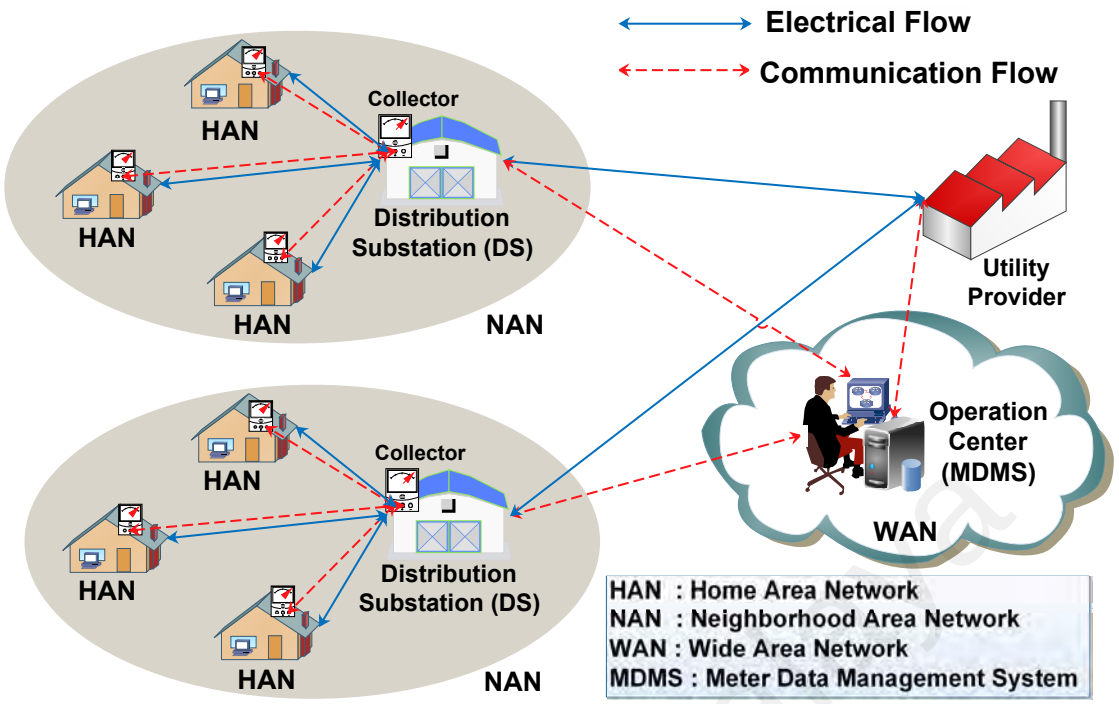



**Figure 2.2: The architecture of AMI in smart grid.**

by the system administrator to transmit data to various parties at predefined time intervals. Since the communication flow in AMI is bi-directional, these IoT devices receive control signals and act accordingly. Besides, the utility pricing information provided by the UPs also allows the load controlling devices to regulate consumption based on consumers' directives and criteria.

In the old days, conventional meters are used only for billing the energy consumed by consumers. In recent years, SMs are introduced along with the deployment of SG to enhance reliability and efficiency of future power systems with DER, as well as distributed demand response (W. Wang & Lu, 2013). SM plays an important role in AMI. It is a digitized energy meter that measures the energy consumption of a consumer regularly at predefined intervals. As compared to conventional meter, it can read fine-grained energy consumption information (i.e., values of voltage, current, frequency and phase angle) and collaborates with the master SM (i.e., data collector) as well as head-end to monitor power quality and securely communicate the data with other parties on a real-time

University of Malaya

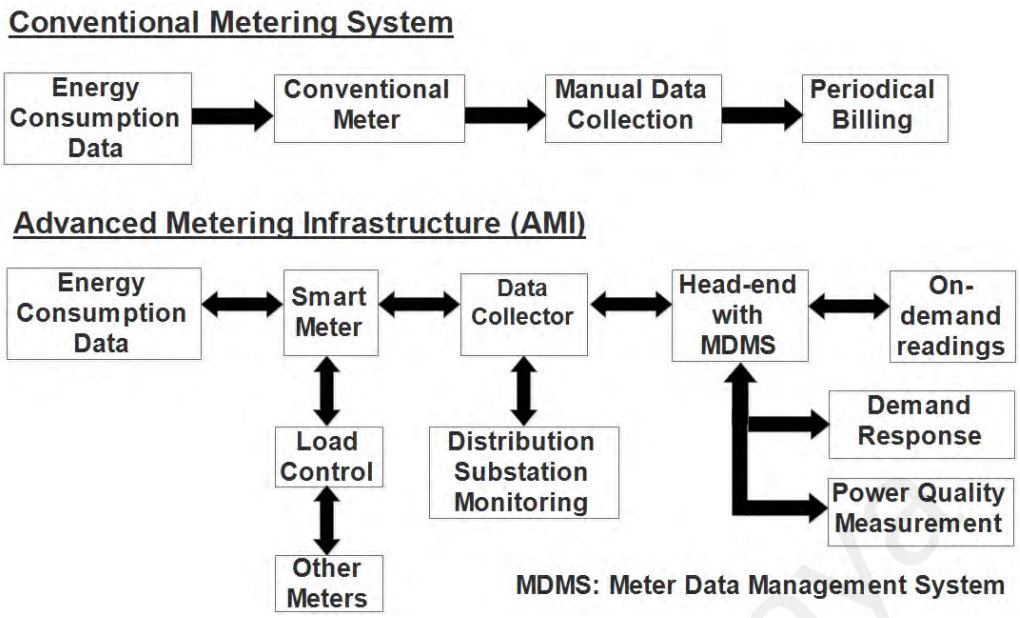



**Figure 2.3: The metering models of conventional power grid and smart grid.**

basis. It can communicate and execute control commands both locally and remotely. Aside from load controlling and monitoring, SM also collects diagnostic information about the home appliances, distribution grid and power related events. SM is also able to restrict the maximum energy consumption by connecting or disconnecting energy supply to any consumer remotely. Data collected by SMs contain parameters such as energy consumption values, time stamp of the data and a unique meter identifier (Depuru, Wang, & Devabhaktuni, 2011). Figure 2.3 depicts the metering models of conventional power grid and SG.

The key features of SM can be summarized as follows:

- Real-time data collection and measurement
- Remote control upgrade/operations
- Outage and failure event notifications
- Time-based pricing
- Power quality monitoring
- Net metering

University of Malaya

- Load scheduling for DSM
- Load limiting for demand response
- Energy theft/meter tampering detection via alarms and sensors
- Efficient energy consumption to improve environmental conditions

#### 2.2.2.2 Communication Infrastructure

As shown in Figure 2.2, the AMI architecture consists of three types of network, namely Home Area Network (HAN), Neighborhood Area Network (NAN) and Wide Area Network (WAN), which are further detailed below:

1. **HAN**: SMs, smart devices, generation, energy storage and plug-in hybrid electric vehicle (PHEV) as well as controllers within the home premises are connected together through HAN. Considering the high data rate and low power transmission requirements, wireless technologies are the optimal solutions for HAN. These technologies include ZigBee, 2.4 GHz WiFi, HomePlug and IEEE 802.11 wireless networking protocol (U.S. Department of Energy, 2010).
2. **NAN**: The data collectors, SMs endowed in consumers' premises and communication networks form a Neighborhood Area Network (NAN). The communication between data collectors and SMs in NAN vary depending on the application scenarios, such as Power Line Carrier, RS485, GPRS/3G and ZigBee.
3. **WAN**: WAN is the transmission network which connects the data collectors to the operation center. Certain networks are adopted for different areas with specific conditions in WAN, such as Power Line Carrier, Ethernet and GPRS/3G.

After data collection, the data are sent to an operation center which contains a MDMS and a head-end. The head-end is responsible for managing communication protocols, collecting and storing metering data, communicating with devices and adapting Internet

University of Malaya

protocols. Meanwhile, the MDMS, which is the central module of the management system, monitors the distribution system. Besides, MDMS is also in charge of data analysis, maintenance and operation of the SGs.

### 2.2.3 Applications and Benefits of AMI

AMI not only offers benefits to consumers, but also to UPs and society (U.S. Department of Energy, 2008).

**Consumers**

- **Reduce electricity billings**: AMI provides a vast amount of grid status and energy usage information. With these information, consumers can make more informed consumption decisions while UPs can make better decisions about service offerings and system improvements. In such a case, AMI provides better power quality, higher reliability and more accurate billing. All these features keep down utility costs, and therefore paying prices for consumers.

**UPs**

UPs' advantages fall into two major categories, operations and billing.

- **Reduce operational and labor costs**: SMs deployed in AMI enables functionalities such as automatic metering data transmission and remote update of SM firmware. Therefore, UPs do not have to send staff to read and maintain meters, thereby reducing operational and labor costs. Besides, unwanted economic losses that are caused by data errors during on-site meter reading can be prevented. Through remote diagnostics, many customer service and maintenance issues can be resolved more cost-effectively and quickly.
- **Facilitate demand response and load control**: AMI allows UPs to continuously monitor aggregate grid load in order to offer consumers price signals that represent

University of Malaya

market prices (i.e., dynamic pricing). It also controls residential consumers' loads by directly manipulating the smart devices during load-constrained periods within consumer-recommended thresholds (i.e., direct load control). Besides, it can interrupt industrial/commercial consumers' loads by directly controlling the utility system operator during seasonal peaks (i.e., interruptible load). With dynamic pricing, consumers are financially motivated to cut down loads during special peak events or during on-peak periods of the day when energy rates are higher (i.e., critical peak pricing). All these features are able to reduce peak demand, which in turn reduces operation of peaking plants, defers the need for new generation, offers relief during capacity-constrained periods and minimizes transmission congestion.

- **Monitor the health of the grid**: The operation center closely monitors line losses by area or time (i.e., month, quarter or year). The real-time monitoring and measurement give UPs an overview of the line losses over the entire service area, thereby allowing them to optimize the whole electrical power infrastructure.
- **Improve outage management system**: With smart metering, UPs can detect and locate outage/failures events more promptly and accurately. Therefore, the repair crews can be dispatched in a more timely and efficient way.
- **Curtail NTLs**: The SMs, data collector and head-end work jointly to analyze and detect malicious attempts to steal energy. AMI is able to detect skeptical thefts on the basis of historical data curves and metering reports, thereby reducing losses incurred due to NTLs.

**Society**

- **Promote a greener future**: AMI is able to produce a greener environment by improving efficiency in energy delivery and utilization. It can accelerate the use

University of Malaya

of distributed generation (DG), which also directly encourage the use of DER. According to Siddiqui (2008), AMI-enabled distribution system would reduce carbon dioxide emissions by up to 25%.

### 2.2.4 Security Issues and Challenges in AMI

In recent years, millions of homes and business premises in the United State alone have been upgraded to SMs. In 2015, Tenaga Nasional Berhad (TNB) Malaysia has commenced a RM2 million pilot project on SG development for more efficient energy supply and delivery. 1,000 SMs were installed in both Malacca and Putrajaya throughout the one-year period of the pilot project (Tenaga Nasional Berhad, 2017).

Although the motivation of SG involves HAN and energy management, the implementations of AMI evolve around the installation of SMs. Subsequently, security issues and challenges associated with AMI grow substantially as the number of SMs increase drastically. In this thesis, security issues resulting from the implementation of the new power infrastructure are looked into. Besides, initiatives which might help mitigating susceptibility to these harmful impacts are also investigated.

#### 2.2.4.1 Consumers' Privacy

In SG, consumers are financially motivated to collaborate with the UPs to control their energy consumption. However, their privacy is violated as they are required to share their fine-grained information. Specifically, third parties can conduct consumers' load profiling with high accuracy by simply analyzing their fine-grained SM readings. These load profiles not only reveal the types of electrical appliance used in consumers' premises but also the number of residents and their daily routines even in the presence of alarm system. Murrill, Liu, and Thompson II (2012) demonstrated that most of the appliances in a premise can be identified by analyzing only a 15-minute interval energy consumption data. These

University of Malaya

information are very valuable to third parties. For example, burglars can choose the most vulnerable target by studying house owners' daily routines and alarm information. On the other hand, advertising companies may conduct unsolicited marketing based on consumers' daily routines and appliances information while insurance firms may increase consumers' premiums based on information extracted from load-profiles (S. A. Salinas & Li, 2015). As for industrial consumers, their load profiles may contain proprietary information about logistics and equipment used. These information provide a competitive advantage to other companies that intend to gain insight into the processes or imitate the industrial operations.

All the aforementioned privacy issues have eroded consumers' trust in the acceptance of SG as the safety and privacy of their data are not protected. In view of the privacy issues, the Dutch Parliament has rejected the implementation of SG in 2009 (Cuijpers & Koops, 2012). Therefore, a secure infrastructure to protect consumers' privacy is imperative in establishing consumers' trust in adopting SG. To safeguard consumers' from privacy invasion, the government needs to establish a regulatory framework. Particularly, privacy regulations and policies should identify the rules for how consumers' data are collected, to what parties the data are shared and the consequences of information abuse. In addition, academia, government as well as energy industry must assess the security and reliability of the smart devices in the laboratory and field more extensively.

#### 2.2.4.2 The Billion-dollar Bug

Consumers' theft in the electrical power distribution system is not something new. According to Northeast Group (2017), the latest estimates indicate that energy theft costs a staggering of $ 96 billions per year globally, with well over half of that happening in the world's emerging markets including India, Brazil and Russia. Since the early days of energy billings, malicious consumers could attempt all kinds of methods such as physical meter tampering to impede the energy flow calculation. The introduction of SM will

University of Malaya

definitely change the nature of energy theft. Specifically, the attacks would change from crude physical system tampering to remote penetration and manipulation of *smart* devices. These sophisticated attacks not only allow consumers to make subtle changes to their energy consumption readings but also terrorists to mount large-scale attacks either on local or national critical infrastructure (McDaniel & McLaughlin, 2009).

The SMs deployed in AMI are built from easily available software and commodity hardware. However, such a heavy dependence on information networking naturally surrenders the SMs to possible vulnerabilities associated with networking and communication systems such as usage loggers, distributed denial-of-service attacks, meter bots, SM root-kits, malware and viruses (McDaniel & McLaughlin, 2009). These digitized meters are extremely appealing for hackers because the vulnerabilities can be easily monetized. Specifically, a "hack" kit can be used by the energy thief to tamper a SM to reduce energy billings. Once these hack kits are commercialized, each vulnerability would result in a “billion-dollar bug” in the energy industry, whereby the costs incurred would not only be measured in consumers’ theft but also in the prices of replacing millions of malicious meters. Besides, the misuse of SMs could also severely harm the electrical power infrastructure. Specifically, the usage misinformation not only might mislead UPs to make incorrect decisions about the capacity and usage but also blind them to impending problems/attacks.

The future of SG is highly dependent on the policies and regulations of respective governments and UPs. Since these laws would assist UPs, consumers and vendors to evaluate risk, they would significantly encourage the adoption of SG. For a smooth transition to a more environmentally sound and less costly power grid, the security problems that AMI introduces should be anticipated and mitigated.

University of Malaya

## 2.3 Electricity Losses in Electrical Distribution System

Electricity losses are the mismatch between amount of energy supplied and amount of energy reported by the consumers (Nagi, 2009). Estimating energy losses in electrical distribution system is one of the important scopes of distribution system performance. Generally, electricity losses that severely affect the UPs can be attributed into two categories namely: (i) technical losses (TLs) due to the distribution and transformation of energy, i.e., proportional to the squared of electrical current, and (ii) non-technical losses (NTLs) that are associated to energy fraud as well as meter irregularities. Ideally, the energy supplied to the service area should tally with the energy recorded. Nevertheless, in reality, these two amounts never tally as electricity losses happen as an integral result of energy distribution and transmission (Nizar, Dong, & Wang, 2008).

### 2.3.1 TLs

TLs in electrical distribution system refer to electricity losses resulting from the energy dissipation or heating of electrical components (i.e., distribution transformer (DT) windings, lines, cables and other measuring equipment) during energy transmission and distribution (Pedro, 2009). TLs cost consumers higher paying price and contribute to carbon emissions. TLs happen as a direct result of the physical characteristics of the electrical equipment used in electrical distribution system. These losses depend on the voltage and transformation levels and the length of the power lines as well as the design of the power grid. These losses include DT losses (i.e., resistive losses and core losses in the windings), resistive losses at the primary feeders, resistive losses in secondary networks, losses due to loose jump wires, losses due to short circuit and earth fault as well as losses in service mains and energy meters (Benedict, 1992).

Generally, TLs are contributed by three main sources (Congres International des Reseaux Electriques de Distribution, 2017): (i) **Load losses (variable losses)** comprising

University of Malaya

of resistive and reactance loss components in the series impedances of the various system elements, (ii) **No-load losses (fixed losses)** that are independent of the actual load served by the power system (Dortolina & Nadira, 2005) and (iii) **Losses due to network services**.

1. **Load losses**: All conductors such as copper wires in overhead lines/cables, coils in transformers, fuses, switch gear and metering equipment, have internal electrical resistance. These resistances cause the conductors to generate heat when carrying electrical current. Load losses are also known as 'variable losses' as energy losses arising from the dissipation of heat to the environment vary with the current flowing through conductors in electrical distribution system. Specifically, load losses vary in proportion to the squared of the current and conductor resistance. Sometimes, these losses are also referred to as "copper losses", "ohmic losses", "resistive losses" or "Joule losses". Besides, deteriorated conductors and loose connections between network equipment might also be the source of this type of losses, as they can cause the arising of heating spots owing to an increase in the resistance. Generally, load losses contribute approximately two thirds to three quarters of the total power system TLs (Vincenzo, Giordano and Georgios, Papaefthymiou, 2015).
2. **No-load losses**: Some electrical energy is dissipated by network equipment or components (e.g., transformers) as a result of being connected to the network and made energized. The system has losses because it is electrically energized even if no energy is delivered to consumers. These losses are dissipated in the form of noise and heat. Most of the no-load losses are usually due to the transformer core/iron losses resulting from the flow of excitation current. These losses are known as "no-load losses" or "fixed losses" because they are independent of the amount of electrical energy the network supplies (Congres International des Reseaux Electriques de Distribution, 2017). In essence, no-load losses depend on applied voltage. They

University of Malaya

do not change with current. Nonetheless, no-load losses are essentially fixed as the applied voltage is typically stable when the network equipment is energized.

3. **Losses due to network services/measuring devices**: Apart from the equipment responsible for the dissipation of energy as load and no-load losses, other equipment connected to the distribution system may also consume energy. Losses can also occur due to uncontracted consumption of network equipment and measuring devices for protection. For instance, measuring elements and network control installed along electrical lines or meters in consumers' facilities are examples of uncontracted consumption that also contribute to TLs.

In short, TLs in the electrical distribution system are fundamentally dependent on electrical loading, network topology and system voltages. It is possible to control and compute TLs, given the known quantities of loads. However, extensive load and network data are required for higher accuracy in estimating the amount of TLs. A variety of approaches has been designed to compute TLs in electrical distribution system (Nadira, Benchluch, & Dortolina, 2003; Dortolina & Nadira, 2005; Au et al., 2008; Oliveira & Padilha-Feltrin, 2009). The parameters commonly utilized to estimate the TLs in distribution network are loss factor, load factor and load profiles/curves (Au & Tan, 2013). In recent years, the enhancements in ICT and data acquisition also make the verification and computation of TLs easier. For instance, Sahoo, Nikovski, Muso, and Tsuru (2015) have designed a temperature-dependent predictive model which utilizes data from DT and SMs to compute amount of TLs and detect localities of energy theft in a service area.

### 2.3.2 NTLs

The losses that take place independently of TLs in power systems are known as NTLs (Nagi, 2009). Generally, NTLs occur due to external actions to the power system.

University of Malaya

NTLs might also happen due to the conditions and loads that have not been taken into consideration in Technical Loss (TL) computations (Suriyamongkol, 2002). NTLs typically relate to a number of ways to deliberately circumvent the UPs (Nagi, Yap, Tiong, Ahmed, & Mohammad, 2008). It is usually more complicated to measure NTLs. Simply measuring consumers' energy consumption and total supplied energy by the UPs alone cannot distinguish NTLs from TLs. As a result, additional types of measurement are required to estimate the amount of TLs. The localities of NTL activities can be potentially detected from the inconsistency between the total consumption and supply with the inclusion of the estimated TLs (Sahoo et al., 2015).

NTLs might lead to detrimental effects on a number of aspects such as political stability, finance and economy. NTLs, especially energy theft is always closely related to governance indicators. Countries with political instability, ineffective accountability as well as massive grafts and corruptions usually have higher levels of energy theft (Smith, 2004). Corruption occurs when the utility personnel working for the UPs is bribed for allowing life-threatening illegal power connections and falsifying the meter readings. The situation becomes more severe when political leaders intervene in the legislation to ensure that supporters and cronies are not penalized for carrying out NTL activities. UPs in some of the corrupted countries were nearly bankrupt because fraudulent personnel continues to collect bribes due to collusion between the government and energy industry. Besides, high NTLs also threaten the financial sustainability of the UPs. Specifically, energy pilfering will definitely lead to reduction in UPs' energy revenue and therefore they are lacking of investment funds to improve the electrical distribution system and capacity. As a result, the economic growth is hampered by irregular and inadequate energy supply due to energy fraud. Effectively, the costs of NTLs are passed down to the legitimate consumers and government. In other words, the energy consumption of the energy thieves or non-payers is subsidized by the

University of Malaya

benign paying consumers, UPs and even the government. Therefore, the mitigation of NTLs is crucial for electrical distribution system to ensure the efficiency of the distribution network will be improved while the costs for the consumers, UPs and government will be cut down.

The major causes of NTLs are listed as follows (Nizar, Zhao, & Dong, 2006):

- Non-payment by consumers
- Meter irregularities
- Energy theft

#### 2.3.2.1 Non-payment by Consumers

All energy consumers will receive regular bills to pay for the energy they consume. However, not everyone pays. The delinquency in paying utility bills tends to compound upon itself. Consumers are less likely to pay if the bill collection system fails. Consequently, a snowball effect happens, whereby the courts become backlogged while UPs take losses. In such a case, it might result in a range of systemic problems that gradually hinder both the economy and UPs's operations (Krishnaswamy, 1999).

To mitigate the non-payment problems, most of the UPs, including TNB Malaysia cuts power supply to the premises over unpaid bills and restores power after the consumers settle the pending bills (Wafi, Aziz, Rahim, Amirhussain, & Norddin, 2013). Besides, prepayment meters are also introduced in some countries to help consumers avoid getting into debt with UPs. With the prepayment meter, consumers pay for their energy before they consume it (Telles Esteves, Cyrino Oliveira, Antunes, & Souza, 2016).

#### 2.3.2.2 Meter Irregularities

As mentioned earlier, NTLs could also happen due to meter irregularity. It commonly takes place when the meter is unable to record the correct energy consumption. According

University of Malaya

to Tenaga Nasional Berhad (2018), meter irregularities occur due to the factors as follows:

- Incorrect meter reading
- Incorrect application of a meter multiplying current transformer (CT) ratio
- Meter inaccuracy
- Malfunctioning of the meter/equipment breakdown
- Faulty installation
- Cross-connection of installation to different accounts

With the advent of AMI technologies, the UPs can now monitor millions of SMs in real time to comprehensively identify and ameliorate NTL problems. For instance, they can easily compute the under-reporting or over-reporting amount based on the consumer's fine-grained energy consumption record and history. Real-time data through advanced grid sensors and smart metering can provide UPs with an overview of the power grids.

#### 2.3.2.3 Energy Theft

Interestingly, meter irregularities and energy theft account for the major causes for the aforementioned NTLs in Malaysia (Tenaga Nasional Berhad, 2006). NTLs have been a daunting problem for the UPs in both developed and developing countries since the beginning of energy billing. The latest estimates show that energy theft totals a staggering $96 billion per year worldwide (Northeast Group, 2017). In United State alone, $6 billion worth of electrical energy is stolen annually (Karaim, 2015). Meanwhile, UPs in India experience losses of approximately $4.5 billion annually due to energy fraud (Ahmad, Chen, Wang, & Guo, 2018). On the other hand, British Columbia Hydro in Canada suffers $100 million every year due to energy stolen for marijuana grow operations (Meuse, 2016). NTL detection becomes more challenging because energy thieves become increasingly sophisticated in their tactics. Energy fraud is not only crippling UPs around the world, but

University of Malaya

also causing higher paying prices for consumers as well as necessitating costly government subsidies. Therefore, the focus of this thesis is the detection of *energy theft and meter irregularities* in SGs.

Various tactics have been exploited to under-report energy. At the consumer level, the most popular technique is to tamper with meter in order to impede proper meter recording or tap energy from a vacant premise. At the grid level, fraudulent consumers always bypass energy meters by wiring heavily-loaded electrical appliances such as heater and air-conditioner directly to the grid. Sometimes they connect the whole electrical system to the feeder with an illegal DT. Meanwhile, at the utility level, inaccuracy in energy billing can result in losses of profits. These inaccuracies can be either unintentional (i.e., meter irregularities as discussed in Section 2.3.2.2) or intentional (e.g., corrupted utility personnel alters the billing record or meter switching with a vacant unit). To curb NTLs from energy theft, TNB Malaysia has formed a special team to conduct physical checks at meters (Tenaga Nasional Berhad, 2006). To improve the effectiveness, the team has been equipped with more technicians as well the purchase of new transportation and equipment to pursue suspected cases of energy fraud, which naturally leads to high utility costs.

Combating energy theft has been one of the key motivations to deploy AMI. In fact, SMs are designed to identify and reveal tampering attempts. The additional features of these modernized solid-state meters as discussed in Section 2.2.2.1 annihilate some attacks that were common in conventional analog meters. Recall that, TNB also started implementing the AMI pilot roll-out in 2015 with the installation of 1,000 SMs in Malacca and Putrajaya (Tenaga Nasional Berhad, 2017). TNB is currently embarking on providing more digitized and automated services not only to offer more values to the consumers but also features such as equipment failure alarms and magnetic tamper detection for NTL analysis.

University of Malaya

Nevertheless, the deployment of smart metering infrastructure with the addition of smart devices and network communications to the electrical distribution system has also introduced new attack techniques. Specifically, SMs are not fully tamper-proof (McLaughlin et al., 2010). For example, an energy thief can easily obtain the root access of the SM to interrupt the communication so that the automated meter alarms and power-related events never reach the UPs. Moreover, the meter alarms are highly susceptible to false positives and hence UPs have difficult task dealing with enormous data to distinguish the fraudulent consumers from the honest consumers.

According to the energy theft techniques discussed in the literature (McLaughlin et al., 2010, 2013; Jiang et al., 2014; Accenture, 2011; Y. Liu, Zhou, & Hu, 2018; Tellbach & Li, 2018), the existing energy theft activities are classified into three categories:

1. Physical attacks
2. Cyber attacks
3. Data attacks

Note that data attacks could also be realized through threats from the cyber and physical attacks. The consumers' consumption data may be compromised at three different stages, namely while it is being recorded, during transmission to UPs or after it is stored (Xiao, Xiao, & Du, 2013). All these attack techniques are presented in Table 2.1. The information in the table will be utilized in Section 5.5 as an attack model for a complete coverage of threats from various known energy theft techniques. Then, the energy thefts and meter irregularities scenarios are simulated by tampering the benign SM readings in order to evaluate the proposed anomaly detection frameworks in Chapter 6.

University of Malaya

**Table 2.1: Types of attack in both conventional power grid and smart grid**

| |
|---|
| **Physical Attacks** |
| Meter switching with an unit from abandoned, vacated or low-consumption premises |
| Neighborhood power diversion |
| Meter tampering<br>• Meter removal/disconnection<br>• Reverse the meter in the socket<br>• Turn back the number dial of electromechanical meters<br>• Place magnets on electromechanical meters<br>• Insert disc to halt rotating of the coil<br>• Damage the rotating coil of the meter<br>• Deposit a highly viscous fluid to damage meter<br>• Remote control switch/relay to control illegal tapping<br>• Modify CT ratio of the meter to impede energy consumption calculation<br>• Abuse optical port to gain root access to the SM |
| Partial wire bypass of the meter |
| Complete meter bypass by wiring heavily-loaded appliances directly to the grid |
| Direct tapping to the primary voltage grid/distribution feeder with an illegal DT |
| Bribe utility personnel for altering billing records |
| Unlawful calibration and indecorous regulation of meters |
| |
| **Cyber Attacks** |
| Steal credentials to login to meters |
| Hack into the firmware of SMs remotely |
| Tamper with the meter storage for information (e.g., recorded total energy consumption, audit logs and encryption key) |
| Compromise meter readings through network exploitation |
| Intercept the meter communications to alter energy consumption values |
| Flood the bandwidth of NAN |
| Exhaust memory/central processing unit (CPU) |
| Erase logged events |
| Interrupt the radio frequency (RF) communication |
| Inject forged values into communication between UPs and SMs |
| Modify traffic between UPs and SMs |
| Meter spoofing |
| RF jamming |
| Design and inject malware into SMs |
| Pricing attacks by manipulating the predictive pricing curve |
| |
| **Data Attacks** |
| Report zero consumption |
| Stop energy consumption reporting |
| Report negative consumption (act as a DG) |
| Remove high-consumption appliances from measurement |
| Under-report the energy consumption by a fraction |
| Modify appliance load profile to hide heftier loads |

University of Malaya

#### 2.3.2.4 Methods of Energy Theft

Since the beginning of energy billing, a vast number of tactics are exploited by fraudulent consumers to alter the meter and its inputs in order to pilfer energy. The common energy theft methods are generally grouped into two major categories, namely (i) meter tampering whereby energy thieves manipulate the internal structure of metering system and (ii) line tampering whereby fraudulent consumers bypass the energy meter connection. The methods of how the energy thieves can get into a position to manipulate the communication signals (i.e., cyber attacks) is not the focus of this paper. Approaches to circumvent encrypted communications in SGs protocols were recently presented in (Jovanovic & Neves, 2015a). The methods used by the energy thieves to fabricate communication signals were discussed in (McLaughlin et al., 2010, 2013; Jiang et al., 2014; Jovanovic & Neves, 2015b, 2015a). In this thesis, the goal is to identify attacks under the assumption that the energy thieves have successfully compromised the integrity of meter consumption readings. In this section, various common tactics exploited to commit energy theft for LV and high voltage (HV) energy meters, recorded by TNB Malaysia (Nagi, 2009) and Provincial Electricity Authority (PEA) Thailand (Millard & Emmerton, 2009) are detailed as follows.

**Low voltage meters (230V single phase)**

In the past decades, conventional analog meter used by developing countries such as Malaysia, Thailand, Vietnam and Indonesia can be easily tampered by breaking the meter seals and gaining access to the components inside. As discussed in Section 2.2.2.1, SMs are equipped with tamper-detection features and encrypted communication capabilities. However, dependence on these functionalities alone is not adequate to guarantee total protection against cyber adversaries who exploit communication vulnerabilities.

A vast number of methods deployed by energy thieves to steal electricity from both LV

University of Malaya

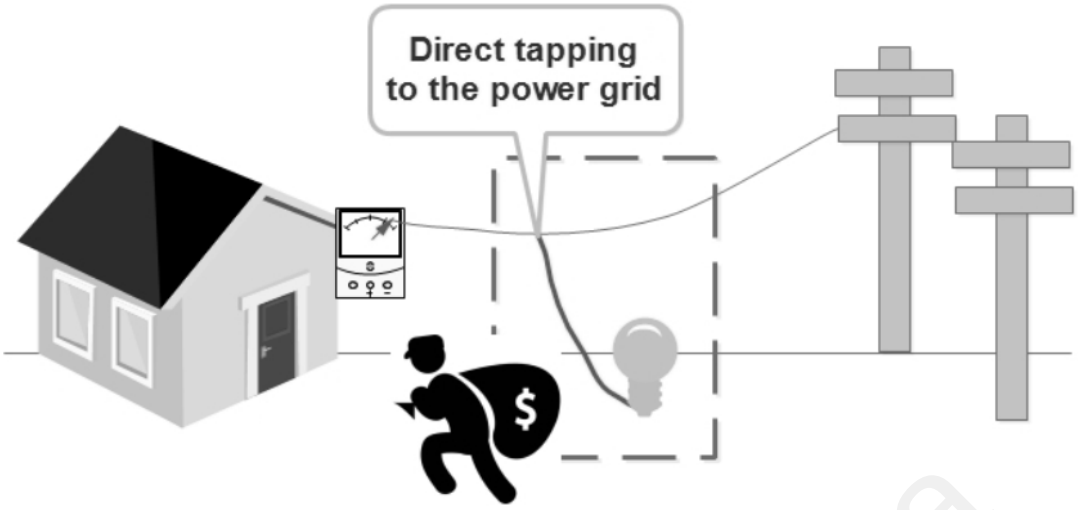



**Figure 2.4: Direct tapping to power line.**

energy meters and SMs are discussed in detail as below.

1. **Directly connecting unregistered load to the power grid (Bypass the meter entirely)**: The energy consumers from the LV network (i.e., domestic or Small and Medium Enterprises (SME)) usually are connected to the LV 230V single-phase/415V three-phase meter and equipment. Therefore, conducting electricity pilfering by direct tapping to the power line is much easier and "safer" as compared to climbing up HV line many stories up on steel masts. Technically, a ladder, a pair of rubber gloves and all the necessary tools are what the energy thieves need in order to perform direct tapping. Direct connection to the power line as shown in Figure 2.4 is the most widely-practised malicious method utilized by many shantytowns and street vendors especially in both Malaysia and Thailand for electricity pilfering.
2. **Meter tampering**: Another popular method of energy theft is by breaking the enclosure seal of the meter. Once the seal of the meter is damaged, energy thieves could easily break into the meter inside the enclosure in order to tamper with the meter. One of the ways is to turn back the number dials in the conventional analog meter to reduce energy billings. However, this method is obviously not applicable for

University of Malaya

digital display meters. Apart from adjusting the number dials, fraudulent consumers might attempt to modify the CT turns of the meter. Besides, the energy thieves could also obstruct the axis and the rotor disk to influence the energy measurements. For instance, fraudulent consumer might insert a thin film to halt the rotation of the rotor disk or deposit highly vicious liquid to damage the meter. Another common form of meter tampering is by placing magnet to slow down the rotation of the disk to confuse the current sensing mechanism, leading to a falsified reduction in measured consumption by 50% to 75% (Evanczuk, 2015). All the aforementioned conventional meter tampering are shown in Figure 2.5. In recent years, electricity poachers are using high-tech equipment and employing elaborated measures to siphon off electricity. Particularly, remote control switch installed in energy meter appears to be the latest preferred gadget used by consumers to steal energy. As shown in Figure 2.6, a remote switching relay is used to bypass the meter to prevent the energy consumption from being registered. With remote control, malicious consumers are able to switch the "bypass switch" on and off whenever they wish. The utility personnel would not be able to identify any irregularities easily when such sophisticated methods, such as remote control switches, are deployed.

3. **Circuit bypass/hidden switch**: In addition, electricity poachers could also steal energy by performing metering circuit bypass simply by connecting jumper wires across the source and load of the meter as shown in Figure 2.7(b). The manner in which meters are tampered with is becoming more and more sophisticated with various advanced gadgets/wrings used to register low readings. A more advanced way to pilfer energy is by installing a "hidden" switch (“Electricity Theft Uncovered at Massage Parlour, Snooker Centre”, 2016). As shown in Figure 2.7(c), the meter will operate normally when switch 1 is closed whenever meters were being monitored

University of Malaya

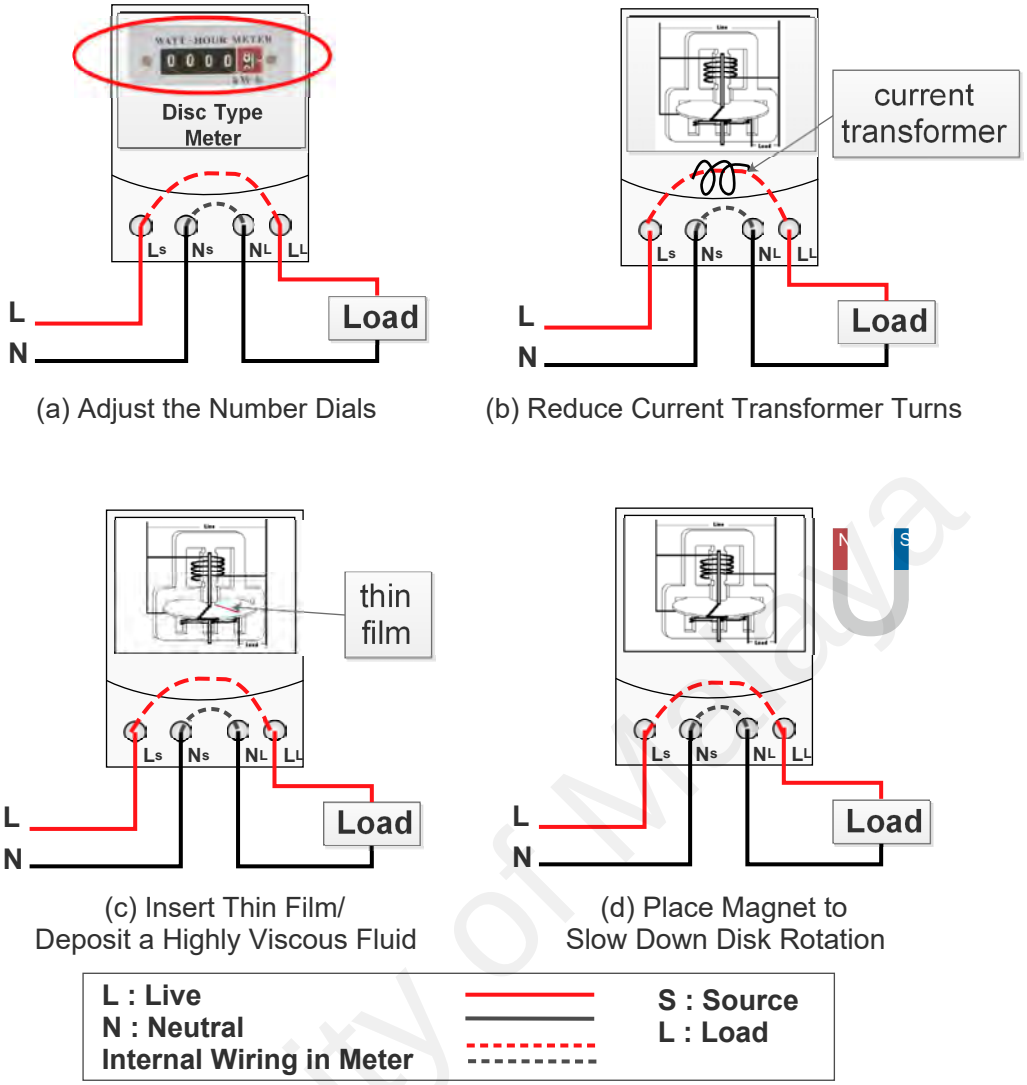



**Figure 2.5: Conventional single-phase analog meter tampering.**

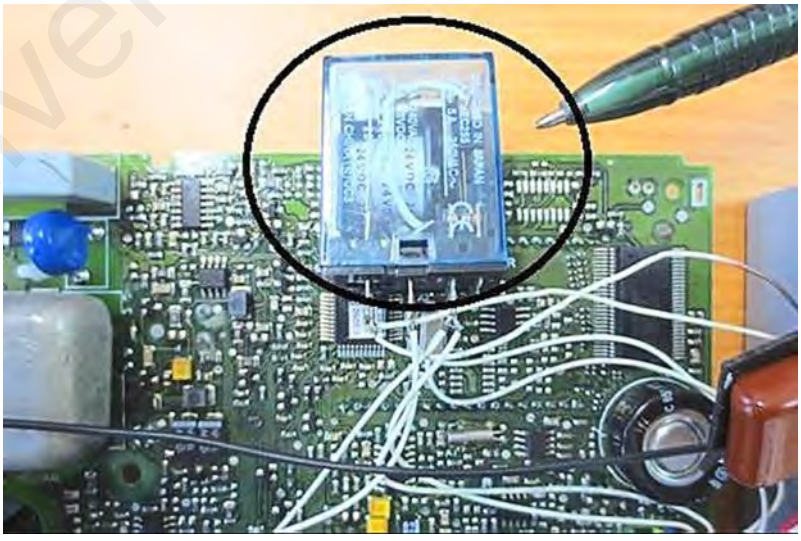


**Figure 2.6: Remote switching relay ("Beauty centre caught stealing electricity using remote control switch", 2014).**

University of Malaya

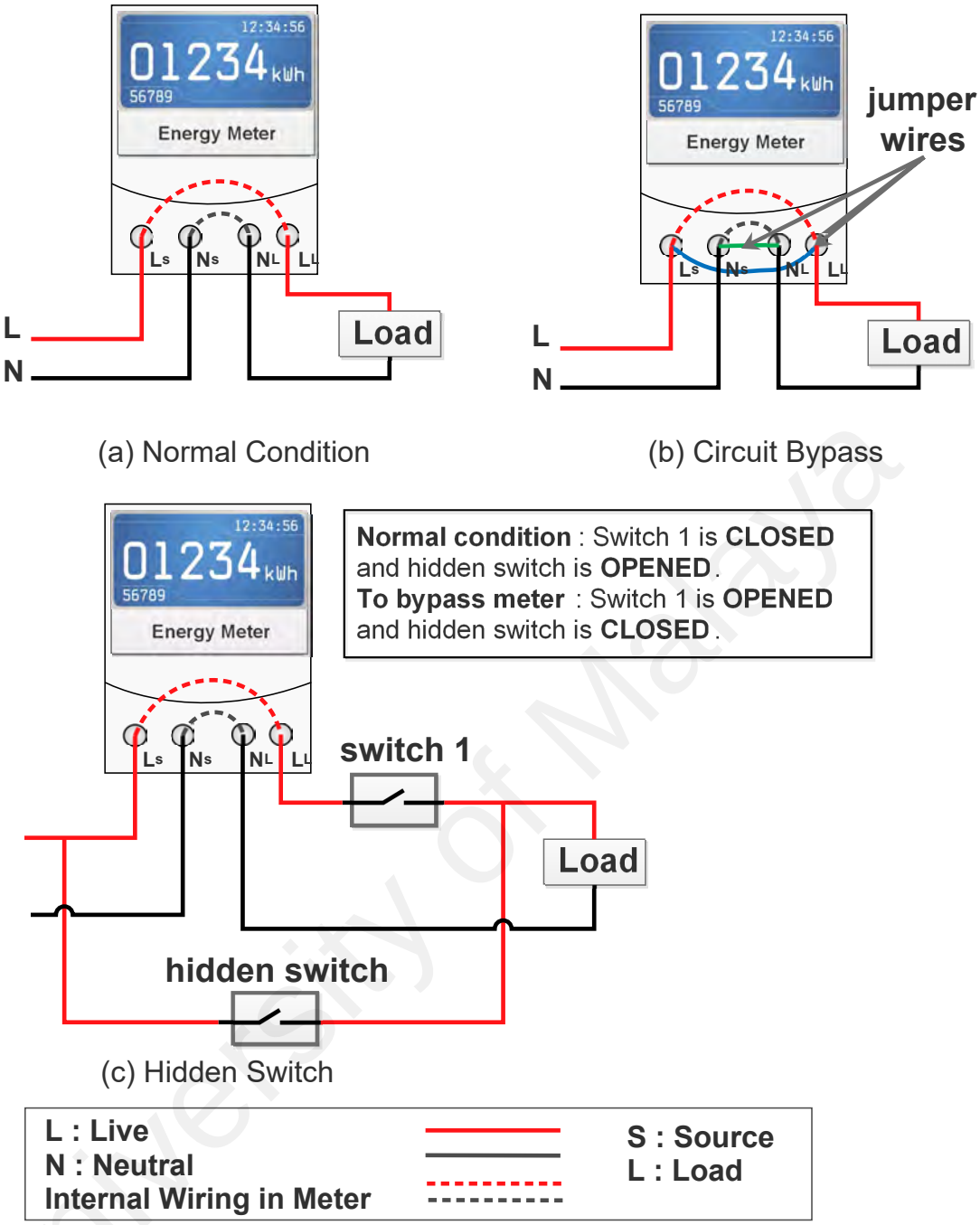



**Figure 2.7: Circuit bypass/hidden switch.**

by UPs. However, the meter will be bypassed when switch 1 is opened while the hidden switch is closed. This type of energy theft becomes even more difficult to be detected if remote control devices are used to control the switches.

University of Malaya

**High voltage meters (12kV or 24kV three-phase, three or four-wire primary)**

Three-phase watt-hour meters are typically endowed in industrial consumers' premises to measure loads that consume high volume of energy and voltage. A technique known as the "two watt-hour meters" connection is utilized in three-phase energy meter to measure energy consumption. Since the loads consume high current and are connected to HV, current and voltage sensing are performed by utilizing CTs and voltage taps, respectively (Suriyamongkol, 2002). According to Tenaga Nasional Berhad (2006), energy theft in commercial and industrial premises contributes the most to NTLs in Malaysia due to the high volume of electricity consumed. Several varieties of tactics exploited by fraudulent consumers to steal energy from HV meters are detailed as below:

1. **Direct connection to the power line**: As compared to LV networks, direct tapping to the HV power line is more difficult as not many electricians would risk themselves exposing to HV power lines without the assistance from UPs.
2. **Meter tampering**: Similar methods that are used to steal energy from LV meters are applied.
3. **Breaking control wires**: Control wires are the secondary wires of the CT. The industrial meters typically measure high currents for large loads. Thus, a step-down CT is connected to reduce current level so that it is compatible with the components in the meter. Electricity pilfering can be accomplished by breaking the insulation of a control wire and connecting external taps to it, thereby causing the meter to under-report energy consumption as the current measured by the meter is reduced. The control wire tampering is shown in Figure 2.8.
4. **Terminal/meter seal tampering**: Since the terminal seals are located below the meter, it is easy to break them. Once the terminal seals are broken, one of the control wires is connected to ground so that at least one phase does not show voltage.

University of Malaya

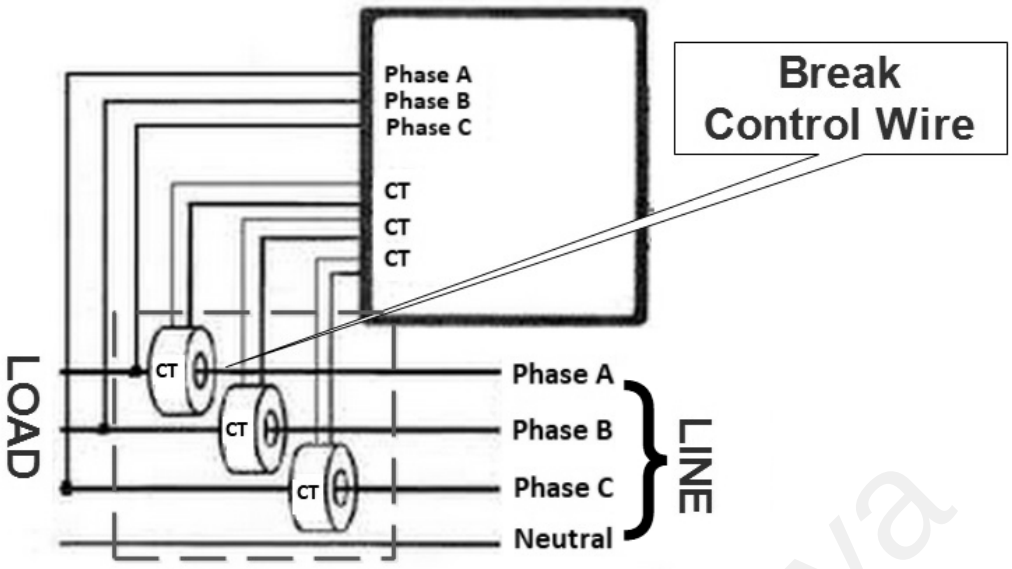



**Figure 2.8: Breaking control wires.**

5. **Breaking voltage taps**: Energy meter reads the voltage of the load by using voltage taps in the meter housing. When the voltage taps are shorted to ground, the meter readings will be distorted (i.e., record lower voltage). Similarly, the meter would record lower voltage if the line is connected to another voltage tap. Nevertheless, most of the meters would not function properly or even would be damaged by voltage tap tampering as the internal components must operate within rated conditions.

## 2.4 NTL Detection Schemes

A variety of energy theft detection strategies (i.e., Support Vector Machine (SVM), load profiling, neural networks, state estimation, decision tree and etc.) (Viegas, Esteves, Melício, Mendes, & Vieira, 2017) have been proposed recently to curb NTL activities. In this thesis, the energy-theft detection schemes are classified into three categories, namely *state-based*, *game theory-based* and *classification*-based schemes (Jiang et al., 2014). However, only data oriented energy theft detection schemes are discussed. Particularly, data oriented schemes are based on data analytics or machine learning. Only consumer related data (e.g., energy consumption measurements and consumer type) are required for

University of Malaya

anomaly detection. Network oriented schemes which utilize power grid data (e.g., network topology and network measurements) for theft detection (Messinis & Hatziargyriou, 2018) are not the focus of this study.

### 2.4.1 State-based Detection

State-based detection scheme leverage on *monitoring state* to detect energy fraud in power system by using specific equipment. The monitoring state can be obtained from mutual inspection (Xiao et al., 2013), wireless sensor networks (McLaughlin et al., 2013), control units (Selvapriya, 2014), radio-frequency identification (RFID) (Khoo & Cheng, 2011) and distribution transformers (Huang, Lo, & Lu, 2013; Sahoo et al., 2015).

Xiao et al. (2013) proposed three inspection schemes to identify anomalous SMs in a neighborhood. Initially, they designed a basic scanning method, which requires linear time to perform inspection. Then, they developed a binary tree-based method for inspection when the *malicious SMs* to *honest users* ratio is high. Finally, an adaptive tree-based method is employed to leverage on the benefits of both the scanning and binary tree schemes. However, a shortcoming of the inspection schemes is the cost. Adding an additional meter for each consumer or UPs will greatly increase the cost on management and hardware. Meanwhile, McLaughlin et al. (2013) designed an AMI Intrusion Detection System (AMIDS). In AMIDS, information of malicious behaviors from three types of information source, namely on-meter anti-tampering sensors, cyber-side network and host-based intrusion detection systems, as well as power measurement-based anomalous consumption detectors are collected, through non-intrusive load monitoring (NILM). However, the adoption of NILM, which requires a high-sampling rate, discloses information about time of use and type of appliance in consumers' premises, hence violating consumers' privacy (Sankar, Raj Rajagopalan, Mohajer, & Vincent Poor, 2013). In (Selvapriya, 2014), consumers' consumption data is compared with the feeder input level. Both individual and

University of Malaya

aggregated consumption are also compared against the feeder details. When the control unit detects any consumption anomalies, the unit alerts the vigilance team by means of Global System for Mobile communications (GSM) message. The utility personnel will then rush to the spot and inspect the premise of the suspected energy fraud. Nevertheless, their proposal can only detect a small region of energy theft but not the exact locality of fraud. Khoo and Cheng (2011) proposed a system that incorporated RFID technology to assist the UPs in ammeter inventory management and ameliorate energy theft. Although RFID technology can be implemented to identify energy , UPs have to pay extra cost to install the system. Thus, the authors adopted cost-benefit theory to examine the value changes caused by the system and then derive a cost-benefit model. Meanwhile, Huang et al. (2013) adopted the measure of overall fit of the estimated values to the pseudo feeder bus injection measurements based on consumers' aggregated meter data at the DT to localize the energy consumption abnormalities. They utilized an analysis of variance to create a list of suspected consumers and estimate the actual consumption based on the state estimation results. Sahoo et al. (2015) designed a temperature dependent predictive model which utilizes SM as well as DT data to estimate TLs and discover fraudulent energy consumption without using the actual system topology information.

### 2.4.2 Game Theory-based Detection

In game theory-based detection schemes, the problem of energy fraud detection is formulated as a game between the fraudulent consumers and the UPs (Cardenas, Amin, Schwartz, Dong, & Sastry, 2012; Amin, Schwartz, Cardenas, & Shankar Sastry, 2015). The goal of the energy thieves is to under-report electricity usage while minimizing the likelihood to be detected. Meanwhile, the UPs intend to maximize the probability of theft detection and minimize the operational cost in managing this anomaly detection mechanism. Game theory-based approaches provide a new perspective into identifying

University of Malaya

and curbing NTLs. The game theory framework proposed by Amin et al. (2015) considers two environments, i.e., perfect competition and unregulated monopoly. A comprehensive game theory model is proposed to analyze the performance of diverse classical statistical techniques for energy theft detection. However, impractical assumptions about the ways fraud is carried out must be made. The studies provide precise detection capacity estimates under the considered assumptions. Cardenas et al. (2012) formulated a game between the energy thieves and UPs. Nash equilibrium of the game is found as the probability density function that defenders and attackers must select before sending AMI measurements. Then, a preliminary analysis of how to choose the maximum sampling interval for SMs in order to safeguard the privacy of consumers while still being able to retain the load shaping attributes of demand response programs. Nevertheless, the formulation of the utility function for all players, i.e., energy thieves and UPs, as well as potential strategies, is still a challenging issue.

### 2.4.3 Classification-based Detection

The key idea of this class of detection techniques is to differentiate abnormal energy consumption patterns from all energy consumption patterns based on a testing dataset containing samples of both the normal and attack classes (Jiang et al., 2014). Several works reported applications of decision tree (Nizar, Dong, Zhao, & Zhang, 2007), Extreme Learning Machine (ELM) (Nizar et al., 2008), rough set (Spirić, Stanković, Dočić, & Popović, 2014), SVM (Nagi et al., 2010) and multi-class SVM (Jokar, Arianpoo, & Leung, 2016) to detect NTLs. Nizar et al. (2007) adopted Naïve Bayesian and decision tree to determine the type of data which provides maximum accuracy with reference to NTL analysis in the electricity distribution sector. In another work, Nizar et al. (2008) investigated the efficiency of SVM technique, ELM and online sequential ELM variant to identify the anomalous consumption trend, which indicates energy fraud based on consumers' load

University of Malaya

profile assessments. Spirić et al. (2014) utilized the rough set theory to identify electricity fraud committed by energy thieves. Based on the amount of uninvoiced/lost electricity due to fraud, they formed a list of suspected consumers. Meanwhile, Nagi et al. (2010) proposed a data mining method together with SVM classifier to detect abnormal behaviors using two-year historical consumption data. The long term trend in energy consumption and computed average daily consumption of consumers were used to detect fraudulent customers. On the other hand, Jokar et al. (2016) adopted multi-class SVM for detecting various types of anomalies in electricity consumption. Aside from the SVM method, other classification techniques, such as fuzzy classification (Dos Angelos, Saavedra, Cortés, & de Souza, 2011) and neural networks (Muniz, Figueiredo, Vellasco, Chavez, & Pacheco, 2009) are adopted to detect energy fraud. Dos Angelos et al. (2011) proposed a fuzzy computational technique to classify energy consumption profiles. Particularly, their proposal consists of two steps. A C-means-based fuzzy clustering is performed to find consumers with similar consumption profiles in the first step. Secondly, fuzzy classification is carried out using the fuzzy membership matrix and Euclidean distance to the cluster centers. Finally, the distance measured are normalized and ordered, resulting in a unitary index score. Therefore, the potential energy thieves with anomalous patterns of energy consumption can be revealed with the highest scores among the index scores. Muniz et al. (2009) proposed an intelligent system to improve the detection accuracy of irregularities among low tension consumers. The proposal consists of two basic modules, namely filtering and classification. Each module is comprised of an ensemble of five neural networks. Then, each network has an output to classify the consumers into irregular or normal. Besides, regression models such as the Auto-Regressive Integrated Moving Average (ARIMA) (Krishna, Iyer, & Sanders, 2016) and Auto-Regressive Moving Average (ARMA) (Mashima & Cárdenas, 2012) have been also adopted for forecasting a time series.

University of Malaya

Assuming that the forecasting model is trained with benign data, the use of ARIMA and ARMA for NTL detection is according to the comparison between forecast and measured values (Messinis & Hatziargyriou, 2018). A larger difference implies higher probability of fraud. Both models are well-known for time series forecasting. Nevertheless, Krishna, Iyer, and Sanders (2016) demonstrated that ARIMA outperforms ARMA for domestic consumers. More recent work by Krishna, Lee, Weaver, Iyer, and Sanders (2016) explored Kullback-Leibler Divergence (KLD) to detect sophisticated electricity theft attacks that circumvent detectors. In their work, KLD is utilized to compare the distribution of a set of measurements with a baseline which is obtained from the historical distribution. The goal of their proposal is to detect a smart attack that disguise anomalous usage as a benign one by fitting it to a legitimate ARIMA model. Thus, it can still identify energy frauds even they are included in the training set. On the other hand, Jindal et al. (2016) designed a decision tree and SVM-based classifiers for rigorous analysis of energy consumption data to detect energy fraud. More specifically, their proposal can be considered as a two-level data processing and analysis approach, since the data processed by decision tree is fed as an input to the SVM classifier. The authors in Villar-Rodriguez, Del Ser, Oregi, Bilbao, and Gil-Lopez (2017) designed a novel algorithm to detect energy consumption outliers in SGs based on concepts from probabilistic data mining and time series analysis. The proposal is able to accommodate time irregularities (i.e., shifts and warps) in the consumption habits of the consumer by concentrating on the trend of the energy consumption rather than on the temporal properties. In a recent work by Buzau, Tejedor-Aguilera, Cruz-Romero, and Gomez-Exposito (2018), the authors proposed an methodology that utilized the auxiliary databases and SM readings to formulate various characteristics of consumers' consumption behavior and also to provide additional information with regard to the geographical and technological characteristics of the SM. These characteristics are then introduced into

University of Malaya

several supervised machine learning algorithms for model selection and evaluation.

Besides, it is crucial to preserve consumers' privacy while detecting energy theft in SGs as detailed in S. Salinas, Li, and Li (2013). In their paper, S. Salinas et al. (2013) proposed a Lower-Upper Decomposition (LUD) algorithm to solve a linear system of equations (LSE) for consumers' honesty coefficients while ensuring consumers' privacy. In LUD, when the consumers steal energy at variable rates, the collector will obtain the honesty coefficient vector and count the number of elements that are not equal to 1. Their proposal can infer by statistics whether it is possible to have that many energy thieves in the community. If it is unlikely for this event to happen, the collector will reduce the sampling period and re-invoke the algorithms again, until the possibility of that event is high and honesty coefficient does not change any more. However, their proposal does not consider technical losses and it is restricted by the dimension of the consumers' energy consumption data (i.e., the data matrix must be a square matrix) due to the characteristic of LUD. In order to meet the dimension requirements, S. Salinas et al. (2013) need to reduce/increase the time granularity. Nevertheless, it might not be practical to *reduce the sampling period* or *increase the time granularity* indefinitely due to the memory size of SM.

However, some of the classification-based detection methods are vulnerable to contamination attacks. Specifically, an energy thief can deceive the learning machine to accept a malicious trend as a normal one through granular changes in data and pollution on the dataset. Besides, most of the machine learning-based detection approaches typically require long term monitoring and measurements before theft detection can be performed accurately. The large sample size requirement generally results in longer detection delay (Jokar et al., 2016). In addition, most NTL detection schemes do not consider TLs, which may prohibit their deployment for actual utilization.

To address some of the limitations of previous work, LR-based and LP-based detection

University of Malaya

frameworks for identifying energy theft and meter irregularities which are not restricted by the dimension of consumers' power consumption data as well as its time granularity are proposed in this thesis. The proposed detection frameworks are able to detect NTL events regardless of whether they occur all the time or at varying rates during intermittent periods in a day. In pursuit of higher detection rate and lower false positives, the impact of TLs and measurement noise on the detection frameworks are taken into consideration. A diverse set of NTL attack functions is investigated and evaluated on the proposed detection frameworks to confirm the reliability of the proposals in real-world AMI energy frauds/metering defects scenarios. No extra hardware costs will incur as UPs can directly apply the proposed frameworks to detect the localities of defective and compromised SMs wholly based on the collected energy consumption data. In such a case, the costs incurred due to NTL events and false positives can be mitigated. This in turn reduces the overall operation costs of UPs and paying prices for consumers.

## 2.5 Summary of Chapter

The goals of all UPs worldwide are to minimize operation costs and maximize revenues, which usually require dealing with electricity losses such as TLs and NTLs. Energy theft, which is the main contributor of NTLs must be ameliorated as these losses contribute to the costs of energy, which is also passed to the benign paying consumers. Fortunately, while the occurrences of NTLs have been increasing, the introduction of SG technologies has brought about better methods to identify and analyze suspected energy fraud. The SMs and smart devices are able to provide fine-grained data that can be leveraged by data analytics and software to detect the localities of energy frauds and metering defects accurately. Leveraging all these smart devices for revenue protection enable UPs to obtain enormous payback benefits from their investment in SG deployment.

From all the existing solutions available which include Supervisory Control and

University of Malaya

Data Acquisition (SCADA) systems, NILM, SVM and on-site investigations, the current approach is to utilize consumers' fine-grained energy consumption data and data analytics as a way of revealing energy fraud and meter irregularities. For this reason, a fraud detection scheme similar to those implemented by energy industry and other businesses, such a bank loan applications and credit card transactions is highly recommended for deployment by UPs.

This chapter studied the fundamental of SGs and various aspects of AMI. The literature and background related to electricity losses in electrical distribution system, including TL and NTL activities were reviewed. In addition, various means of energy theft including direct connections, meter tampering and speed reduction of the rotating disk in both LV and HV energy meters were discussed. In the end, a comprehensive review of existing NTL detection schemes, which include state-based, game-theory-based and classification-based detection, was also presented.

University of Malaya

# CHAPTER 3: LINEAR REGRESSION-BASED ANOMALY DETECTION FRAMEWORK

## 3.1 Overview

This chapter presents in detail two anomaly detection schemes for identifying energy theft and meter irregularities. As discussed in Chapter 2, UPs suffer tremendous losses up to billions of dollars annually due to NTLs (Northeast Group, 2017). Therefore, it is crucial to explore different approaches to mitigate NTLs due to electricity thefts and inaccurate meters readings. To address NTLs, UPs such as British Columbia Hydro are convincing consumers to install SM in their household so that UPs can leverage on the energy consumption data collected from the AMI to identify possible defective meters and abnormal consumers' consumption patterns (Krishna, Lee, et al., 2016). In this chapter, two *linear regression-based* detection schemes are designed to study consumers' energy utilization behavior and their *anomaly coefficients* are evaluated to combat energy theft caused by meter tampering and detect defective SMs. Any anomalies not following the utilization trend may be indicative of energy thefts or metering defects. Categorical variables and *detection coefficients* are also introduced in the framework to identify the periods and localities of energy frauds as well as faulty SMs when NTLs only take place during a certain period in a day.

## 3.2 Motivation

Most state-of-the-art SMs (Comed, 2017) are equipped with tamper-detection and encrypted communication features. Nonetheless, dependence on these security mechanisms alone is inadequate to ensure total defense against cyber-intrusions which exploit communication and network vulnerabilities. Specifically, AMI can be exploited by the energy thieves to perform a number of attacks for falsifying the energy utilization statistics because SMs are vulnerable to more sophisticated types of NTL attack such as

University of Malaya

network-borne attacks. Consumers' consumption data may be compromised at three different stages, namely, during transmission to utility provider (UP), while it is being recorded, or after it is stored (Xiao et al., 2013). Since the conventional methods for mitigating NTLs impose high operational costs (e.g., on-site inspection where extensive deployment of human resources is involved (Nizar, Dong, Jalaluddin, & Raffles, 2006)), this research aims at reducing the operational costs of UP by detecting NTL activities through deployment of AMI in SGs. In general, the existing NTL detection schemes are vulnerable to contamination attacks/non-malicious factors and require large sample size for detection analysis, thereby limiting the detection rate (Jokar et al., 2016). Therefore, an anomaly detection framework that can efficiently detect energy theft attacks against AMI has become significantly imperative for reducing costs and revenue losses incurred due to NTLs. In this chapter, the goal is to detect NTL attacks under the assumption that the energy thieves have successfully compromised the integrity of consumers' consumption readings.

## 3.3 Architecture of Smart Grid in Neighborhood Area Network

This section presents the communication and electrical network architectures considered in this thesis. In AMI, the electrical and communication networks overlay each other and all electrical and communication flows are bidirectional (Fang et al., 2012). According to the surveys of SG (Li et al., 2010; Yan et al., 2013), the architecture of SG in a NAN can be illustrated in Figure 3.1. Further details on *Electrical Network* and *Communication Network* will be provided below.

### 3.3.1 Communication Network

The SMs installed in households, data collector, operation center and DS form a NAN. In a NAN, UP relies on an operation center to monitor the DS and distribution networks. The

University of Malaya

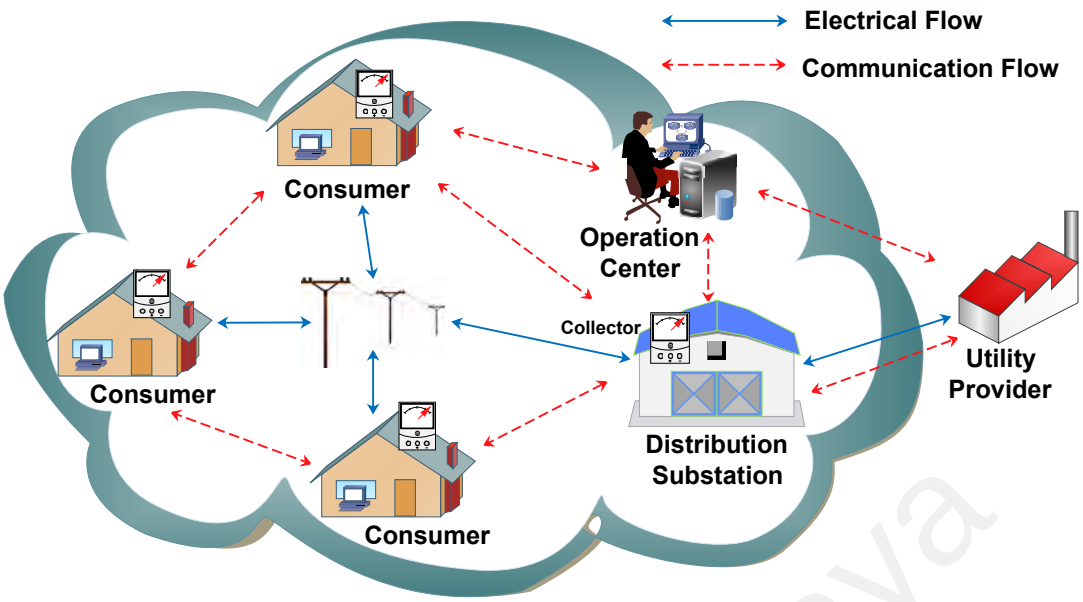



**Figure 3.1: The architecture of smart grid in neighborhood area network.**

communication among the SMs and the data collector are conducted in a wireless manner while the communications among the collector, operation center and DSare conducted via wired medium such as power feeder line (J. Liu, Xiao, & Gao, 2014). In this study, it is assumed that all consumers' premises are endowed with a SM. Therefore, the effect caused by consumers without a SM is not considered.

### 3.3.2 Electrical Network

Similar to the conventional electrical grid system, the power supply of SGs in a NAN is usually serviced by the same UP. The UP builds a DS, which is also known as fuse box (J. Liu et al., 2014) within every neighborhood. The DS acts like an 'electricity router' to distribute power to all the consumers in the neighborhood. A master SM, known as the ***data collector*** is endowed inside the DS to measure the aggregated power supply from the UP to all consumers at a utility selected interval (e.g., every 50 consumers per phase) in the NAN (Accenture, 2011) at time interval $t_i$, denoted by $c_{t_i}$, but not the power consumption of each consumer.

Therefore, in order to track the power consumption of each consumer $n \in \mathbf{N} =$

University of Malaya

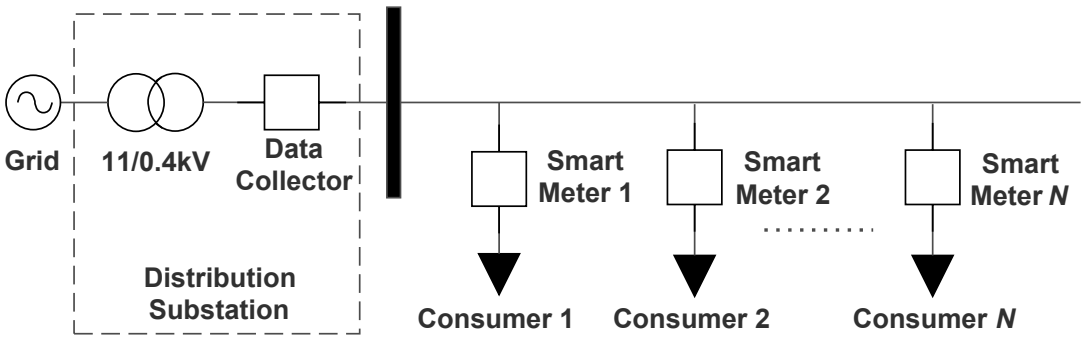



**Figure 3.2: A radial electrical network topology in neighborhood area network.**

$\{1, 2, \cdots N\}$, UP installs a SM at each consumer's household. The SM of consumer $n$ automatically records energy consumption as a function of time interval $t_i$ (subject to the time granularity of the SM), denoted by $p_{t_i,n}$ and computes the consumption cost of each household. Specifically, the SM reading is recorded at time stamp $t_i$, where the interval is $t_i - t_{i-1}$ (Han & Xiao, 2014).

Since most power distribution networks are radial in practice, only radial topology as shown in Figure 3.2 is considered. In the NAN as illustrated in Figure 3.3, the leaf nodes (i.e., consumers) are connected to the root node (i.e., DS). Besides, the losses (i.e., both TLs ($\lambda$) and NTLs due to energy theft ($\theta$) and faulty SMs ($\gamma$)) are also modeled as leaf nodes. Since active power is additive, the total energy supplied by the root node to the NAN should tally with the sum of electricity consumption reported by all the leaf nodes at time interval ($t_i$) (Rashed Mohassel et al., 2014). Thus, the following equation holds:

$$c_{t_i} = \sum_{n=1}^{N} p_{t_i,n} + \lambda + \theta + \gamma, \tag{3.1}$$

where $\lambda$ denotes the TLs, while $\theta$ and $\gamma$ indicate inaccurate meter readings due to energy thefts and faulty SMs, respectively.

Therefore, if $\theta > 0$ (i.e., energy theft exists) or $\gamma < 0$ (i.e., at least one SM is malfunctioning), the discrepancy in meter reading at time interval $t_i$, denoted by $y_{t_i}$, is

University of Malaya

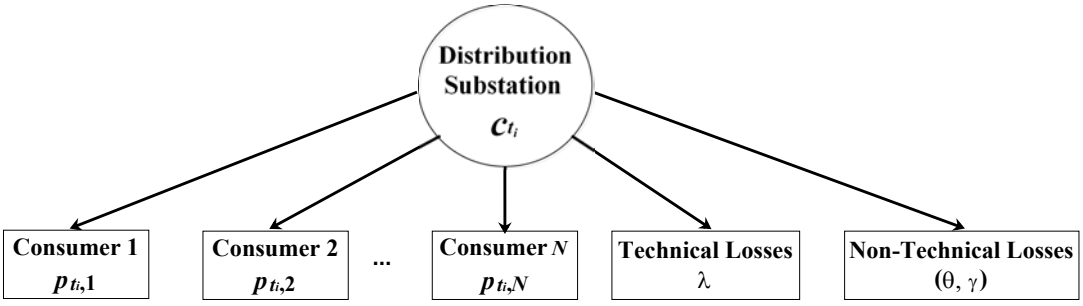



**Figure 3.3: Illustration of a radial electrical network topology. Circle represents the root node (i.e., distribution substation). Rectangles represent the leaf nodes (i.e., consumers and electricity losses).**

computed as:

$$y_{t_i} = c_{t_i} - \sum_{n=1}^{N} p_{t_i,n} = \lambda + \theta + \gamma. \tag{3.2}$$

## 3.4 LR Model for Detecting NTLs

In this section, the mathematical model for detecting energy theft and meter irregularities in a NAN is presented. Suppose that UP equips a SM at each household to record the electricity consumption at some predefined time intervals. Meanwhile, a data collector is installed inside the DS such that it can measure the aggregated power supply from the UP to the service area.

Consider a service area consisting of $N$ consumers. Recall that $p_{t_i,n}$ and $c_{t_i}$ denote the near real-time energy consumption recorded by consumer $n$ and data collector, respectively, at time interval $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$. Then, an *anomaly coefficient*, denoted by $a_n$ is further defined for each consumer such that $a_n = 0$ if consumer $n$ is honest in reporting his energy consumption. Therefore, $(a_n + 1)p_{t_i,n}$ gives the cumulative energy consumption reported by consumer $n$ at $t_i$. As discussed in Section 3.3.2, the sum of electricity consumption reported by all the consumers must agree with the total load consumption measured by the collector at time interval $t_i$ (S. Salinas et al., 2013), the following equation

University of Malaya

can be formulated:

$$(a_1 + 1)p_{t_i,1} + (a_2 + 1)p_{t_i,2} + \ldots + (a_n + 1)p_{t_i,n} = c_{t_i}. \tag{3.3}$$

Re-arranging which gives:

$$a_1 p_{t_i,1} + a_2 p_{t_i,2} + \ldots + a_n p_{t_i,n} = c_{t_i} - \sum_{n=1}^{N} p_{t_i,n}. \tag{3.4}$$

Similar to Equation (3.2), the right hand side of Equation (3.4) is the difference between the total electricity supplied by the UP and the sum of energy consumption reported by all consumers in the service area at time interval $t_i$.

Note that the proposed LR model does not consider TLs (in which its percentage is denoted by $\lambda$) in the SGs. TLs occur during power distribution and transmission, which involve DS, transformers and line-related losses (Nizar, Zhao, & Dong, 2006). TLs might also occur due to dynamic environment factors (e.g., temperature) and are caused by the LV power lines as well as intrinsic inefficiencies in the transformers (S. Salinas et al., 2013). Nonetheless, Sahoo et al. (2015) proposed a method to precisely compute TLs in branches of distribution system. In their proposal, a specific circuit is assumed for each branch. By applying the least square regression to the data from distribution transformers and the current readings collected by smart or conventional power meters, the resistances of the lines connecting the consumption points to the distribution transformers as well as the non-ohmic losses are calculated. These parameters are then utilized to predict TLs in future time intervals. Thus, once the TLs are calculated from Sahoo's approach, the proposed model can be adjusted accordingly by subtracting TLs from vector $\mathbf{y}$ as expressed in Equation (3.2).

The main goal is to find anomaly coefficient of consumer $n$ ($a_n$) in the system of equations

University of Malaya

**Table 3.1: Description of $a_n$**

| Scenario | Description |
|---|---|
| $a_n = 0$ | Consumer $n$ is honest in energy consumption reporting |
| $a_n > 0$ | Consumer $n$ under-reports what was consumed |
| $a_n < 0$ | The $n$-th SM over-reports what was consumed |

from Equation (3.4) in order to evaluate the anomalous behavior of each consumer or reliability of SM endowed in each household. In particular, there are three possibilities as summarized in Table 3.1.

Suppose that the electricity consumption is sampled over $T$ time intervals in a day. A LSE for the detection of electricity theft and metering defects can be formulated as follows:

$$\begin{cases} a_1 p_{t_1,1} + a_2 p_{t_1,2} + \ldots + a_N p_{t_1,N} = y_{t_1} \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_1 p_{t_T,1} + a_2 p_{t_T,2} + \ldots + a_N p_{t_T,N} = y_{t_T} \end{cases} \tag{3.5}$$

The LSE can also be expressed in matrix-vector form:

$$\mathbf{Pa} = \mathbf{y} \tag{3.6}$$

where

$$\mathbf{P} = \begin{bmatrix} p_{t_1,1} & p_{t_1,2} & \cdots & p_{t_1,N} \\ p_{t_2,1} & p_{t_2,2} & \cdots & p_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{t_T,1} & p_{t_T,2} & \cdots & p_{t_T,N} \end{bmatrix},$$

$$\mathbf{a} = [a_1, a_2, \cdots, a_N]' \text{ and } \mathbf{y} = [y_{t_1}, y_{t_2}, \cdots, y_{t_T}]'. \tag{3.7}$$

Here, the $t_i$-th row of $\mathbf{P}$ represents the data recorded by all $N$ consumers at the $t_i$-th

University of Malaya

time interval. On the other hand, the $n$-th column of **P** denotes the data measured by the SM for consumer $n$ over all $t_i$. In this model, **a** is a column vector consisting of anomaly coefficients $a_1, a_2, \cdots, a_N$.

The scenario is explained using a simple 2-consumer topology, namely consumer $A$ and consumer $B$, respectively. As mentioned previously, if there are no energy thefts or defective SMs at $t_i$, meter discrepancy at time interval $t_i$ ($y_{t_i}$) is equal to 0 in Equation (3.2). Then, Equation (3.4) becomes $a_A p_{t_i,A} + a_B p_{t_i,B} = y_{t_i} = 0$ because the sum of consumption readings of all consumers matches the total power supplied by the UP. In particular, both $a_A$ and $a_B$ are 0 as the energy reporting of the consumers are truthful. However, $y_{t_i} \neq 0$ implies that either the AMI is under attack or one or more of the SMs may be faulty at $t_i$. If consumer $A$ is honest while consumer $B$ reports less than what was consumed, then $a_A = 0$ and $a_B > 0$. Similarly, $a_A > 0$ and $a_B = 0$ happen when consumer $A$ cheats on the SM readings while consumer $B$ is honest.

## 3.5 Estimating Anomaly Coefficients using Linear Regression

In the following sections, two schemes are developed to solve the LSE for the anomaly coefficients in Equation (3.6) using *Linear Regression (LR)*. The objective is to enable the data collector to reveal the localities of energy thieves and/or faulty SMs.

### 3.5.1 Multiple Linear Regression

A **L**inear **R**egression**-**based scheme for Detection of **E**nergy **T**heft and **D**efective Smart **M**eters, hereafter referred to as **LR-ETDM**, is first developed to detect energy thieves and defective SMs. MLR is a modeling technique utilized to explicitly describe the relationship between a continuous-valued response $Y_i$ and linear predictors $p_{t_i,1}, p_{t_i,2}, \cdots, p_{t_i,N}$. The goal of regression analysis is to find a function that describes, as closely as possible, the relationship between the variables so that the value of the dependent variables can

University of Malaya

be estimated using a range of independent variables (Amral, Ozveren, & King, 2007; Schneider, Hommel, & Blettner, 2010). Here, $y_{t_i}$ as defined in Equation (3.2) is viewed as the realization of a normally distributed random variable $Y_i \sim N(d_{t_i}, \sigma^2)$, where

$$d_{t_i} = \alpha + \sum_{n=1}^{N} a_n p_{t_i,n}. \tag{3.8}$$

Equation (3.8) defines a hyper-plane (Rodriguez, 2013), where the parameter $\alpha$ (i.e., known as intercept) represents the expected response when all the predictors are zero, i.e., $p_{t_1,1} = \cdots = p_{t_i,n} = 0$. The parameter $a_n$ represents the expected increment in the response per unit change in $p_{t_i,n}$ when the other predictors are constant. In this study, $\alpha$ is set as 0 due to the assumption that the response is entirely dependent on the predictors.

An important characteristic of the linear regression-based model (i.e., Equation (3.8)) is that it is additive (Rodriguez, 2013). Specifically, the effect of a predictor on the response is always the same regardless of the values of the other predictors. The implicit assumptions are:

1. **The predictors are uncorrelated with each other**. In other words, there is no linear dependencies among the predictors. This assumption is reasonable so it does not warrant changes to the model as expressed in Equation (3.8).
2. **The coefficients $a_n$ never change throughout the period of observation**. This assumption only holds true when the consumers cheat consistently throughout the period of observation.

However, inconsistent cheating in energy reporting will lead to inaccurate energy fraud and metering defects detection. Hence, it is possible for some of the dishonest consumers to escape detection when their cheating behaviors change during the period of observation. In this section, it is assumed that consumers steal energy or SMs are damaged all the time.

University of Malaya

This assumption may be unfeasible, and therefore later in Section 3.6, an enhanced model which captures the changes of the estimated anomaly coefficients to identify the period of energy fraud and/or metering defects will be introduced.

It has been shown in (Rodriguez, 2013) that the maximum likelihood estimate of the coefficients **a** are those that minimize the residual sum of squares between $y_{t_i}$ and $d_{t_i}$. If **P** is of full column rank, then **a** is given by:

$$\mathbf{a} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{y}. \tag{3.9}$$

### 3.5.2 Student's *t*-statistic and Two-tailed *p*-value Approach

As mentioned in the previous section, Equation (3.9) is introduced to compute the absolute value of all anomaly coefficients, **a**. However, there is no objective way to determine whether the value of the computed anomaly coefficient is 0 or 1. In LR, the purpose of *t*-statistic is to make inferences about each estimated anomaly coefficient $a_n$ to test the null hypotheses that it is equal to zero. In other words, it means that $a_n$ is likely to be 0 if its corresponding *t*-statistic is not significant, and vice versa.

For a hypothesis test on coefficient $a_n$, with

$$\begin{cases} H_0 \ : \ a_n = 0 \\ H_1 \ : \ a_n \neq 0 \end{cases}, \tag{3.10}$$

the *t*-statistic for estimated $a_n$ is computed as $t = \frac{a_n}{SE(a_n)}$, which follows a *t*-distribution with (*m* - *p*) degrees of freedom (Rodriguez, 2013; Studenmund, 2014). $SE(a_n)$ is the standard error of the estimated anomaly coefficient $a_n$, *m* denotes the number of observations and *p* is the number of regression coefficients.

Each *t*-statistic tests for the significance of each $a_n$ given other coefficients in the

University of Malaya

model. Meanwhile, $p$-value is a function of the $t$-statistic that is utilized for comparing the probability of rejecting $H_0$ when it is actually true. The $p$-value will be compared against a threshold value, known as the significance level, under a *two-tailed test*. The significance level of 5% or 1% are conventionally used as the cut-off between significant and non-significant results (Artes, 1997), but in this study, the latter is chosen to reduce the rate of false positives. If the $p$-value is smaller than a 1% significance level, it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true and hence, the null hypothesis $a_n = 0$ must be rejected. It also implies that there is a relationship between the independent variable and the dependent variable. In other words, it indicates that the anomaly coefficient of consumer $n$, i.e., $a_n$, significantly contributes to the value of the dependent variable (i.e., $y_{t_i}$) in the model.

### 3.5.3 LR-ETDM

In this section, the Linear Regression-based scheme for Detection of Energy Theft and Defective Smart Meters (LR-ETDM) scheme is detailed. Here, a constant scenario is assumed where the fraudulent consumers always steal energy and the defective SMs always report more than what the corresponding consumers actually consumed.

The flow chart as shown in Figure 3.4 summarizes the LR-ETDM scheme. Assume that the data collector labels the SM of all consumers in the service area of interest from 1 to $N$. The $n$-th SM then transmits $p_{t_i,n}$ to the collector to allow the collector to collaboratively compute $y_{t_i}$, $a_n$, $t$-statistic and the corresponding $p$-value. The scheme commences by computing the discrepancies between the total power supplied by the UP (i.e., measurement of data collector at time interval $t_i$ ($c_{t_i}$)) and the total energy consumption of all consumers in the service area (i.e., $\sum_{n=1}^{N} p_{t_i,n}$) for time interval $t_i \in \mathbf{T}$. Then, a LSE consisting of consumers' reported load data, anomaly coefficients and the differences in reading is formed as expressed by Equation (3.5). In this work, the `fitlm` function packaged in the

University of Malaya

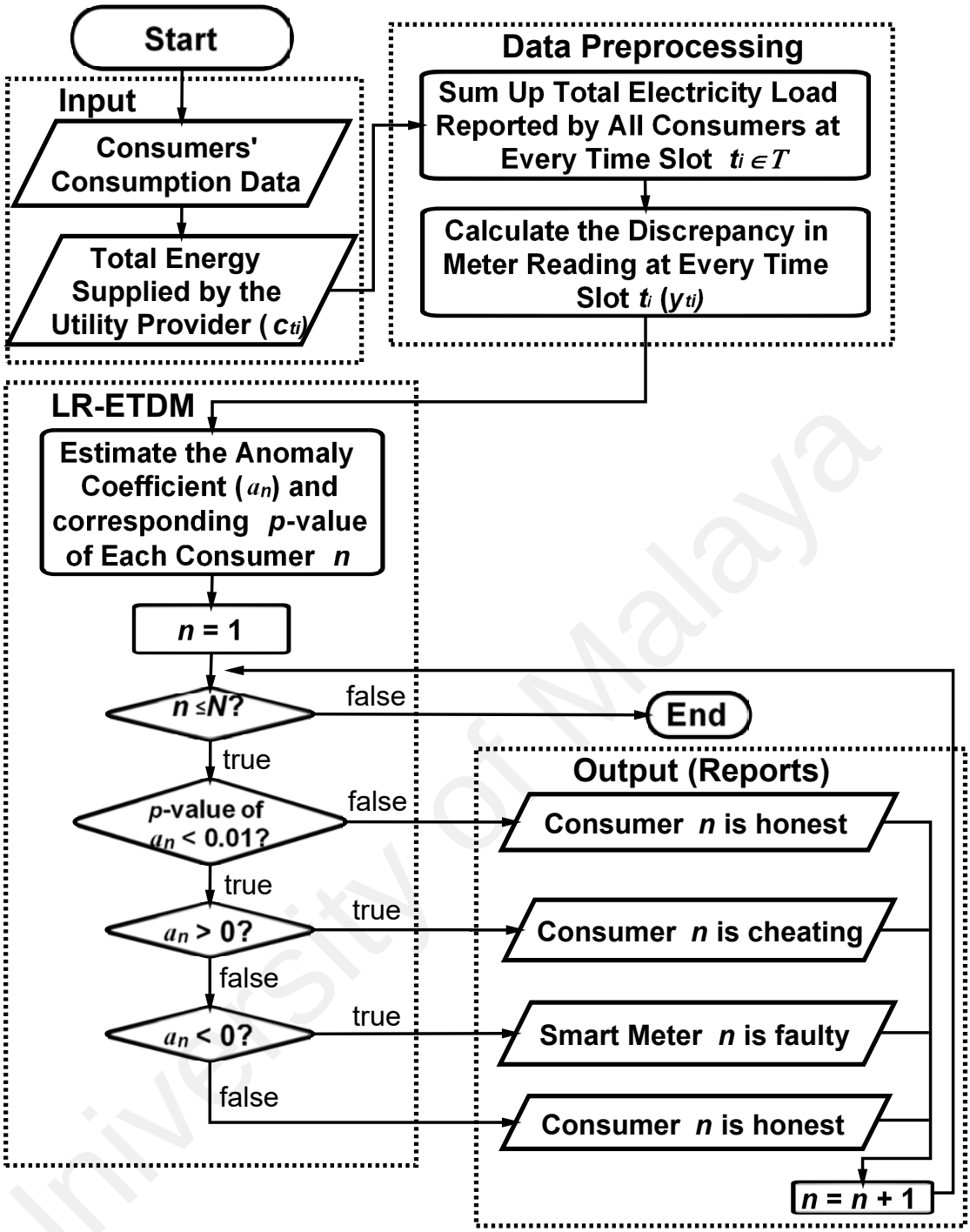



**Figure 3.4: Flow chart of the LR-ETDM scheme.**

Statistics Toolbox of Matlab R2014b is used to solve for the estimated anomaly coefficients $a_n$, standard errors, $t$-statistics and $p$-values. The indicator for the constant intercept in the fit (i.e., $\alpha$ in Equation (3.8)) is configured as 'false' so that the response is entirely dependent on the predictors **P**. Next, the $a_n$, $t$-statistics and corresponding $p$-values of all consumers (i.e., $\forall n \in \mathbf{N}$) are found using LR method. Based on the estimated $a_n$ and

University of Malaya

$p$-values, the locations of energy frauds and faulty SMs can be pinpointed accurately.

For every consumer $n \in \mathbf{N}$, if the $p$-value of the $t$-statistic of consumer $n$ is less than 0.01, it is obvious that this coefficient is significant at the 1% significance level given the other estimated anomaly coefficients in the model, and hence the null hypothesis $a_n = 0$ will be rejected. Specifically, when an energy fraud or metering defect has occurred at household $n$, it is unlikely that $a_n = 0$. In such a case, the estimated anomaly coefficient of the consumer $n$ is further investigated. Obviously, if the predicted $a_n > 0$, it means that the consumer $n$ is reporting less than what he/she consumes. On the contrary, $a_n < 0$ indicates that the SM of consumer $n$ is reporting more than what he/she consumes. In other words, the SM may be malfunctioning. Otherwise, if $a_n = 0$ or $p$-value of $a_n > 0.01$, consumer $n$ is honest and hence the SM is neither fraudulent nor faulty. Note that the collector invokes LR-ETDM scheme at the end of each day after data collection has completed.

It is observed that LR-ETDM may not be numerically stable when the fraudulent consumers do not steal energy constantly. Specifically, LR-ETDM may not detect all energy thieves when fraudulent consumers only cheat during a particular period in a day. For instance, they only cheat during the peak hours. The inaccuracies are due to the limiting factors of regression model. As discussed in Section 3.5.1, linear regression explicitly assumes that the anomaly coefficients $a_n$ do not change throughout the period of observation (Chambers & Dinsmore, 2014). In other words, linear regression presumes that if a consumer cheats, he/she cheats at the same rate throughout the day. Thus, some of the dishonest consumers could stay undetected when they do not cheat all the time.

Therefore, in Section 3.6, the assumption of constant anomaly coefficients is removed and an enhanced scheme is put forward to reveal the locations and periods (i.e., during off-peak, on-peak of a day or whole day) of energy theft or device failure by introducing *categorical variables* in linear regression.

University of Malaya

## 3.6 Estimating Varying Anomaly Coefficients using Categorical Variables

In LR-ETDM, it is assumed that the anomaly coefficients, $a_1, a_2, \cdots, a_N$ are constant. However, it is possible that the rate at which the fraudulent consumers steal electricity is variable when they commit energy theft (S. Salinas et al., 2013). In SGs, time-of-use (TOU) pricing scheme is also present in AMI. TOU scheme refers to a pricing scheme in which energy costs more during peak load period, and vice versa. Specifically, TOU scheme divides a day into several periods known as tariffs, typically off-peak and on-peak (McLaughlin et al., 2010) tariffs. Therefore, consumers will be motivated to reduce energy costs by shifting some energy-intensive loads to off-peak hours or tampering with the SM readings during the peak demand period. It is observed that when dishonest consumers attempt to falsify their energy consumption inconsistently, LR-ETDM gives an anomaly coefficient vector where some of the predicted elements are showing inaccurate values. To overcome the deficiency in the LR-ETDM scheme, another scheme, **C**ategorical **V**ariable-Enhanced **L**inear **R**egression-based scheme for detection of **E**nergy **T**heft and **D**efective Smart **M**eters (**CVLR-ETDM**) is proposed, by introducing *categorical variables* in linear regression through *dummy coding* to resolve the varying cheating problem.

### 3.6.1 Categorical Variables in Regression: Dummy Coding

Linear regression allows the inclusion of categorical independent variables known as dummy variables through *dummy coding*. It is utilized when one wants to compare other groups of the predictor variables with one specific group of predictor variables (i.e., reference group) (Pedhazur, 1997). Dummy variables take the values of 0 or 1. Specifically, the value of 0 and 1 imply the absence and presence of the attribute of the category, respectively. It is necessary to create $k - 1$ dummy variables where $k$ indicates the number of categories of the predictor (Starkweather, 1997; Skrivanek, 2009).

In this study, the categorical variables, $x_i$ for $i = 1, 2, \cdots, N$ are included to categorize



University of Malaya

the time of fraud or metering defect of consumers $1, 2, \cdots, N$. The period of energy theft or metering defect is grouped into two categories, namely off-peak (i.e., from 08:00 P.M. to 07:59 A.M.) and on-peak (i.e., from 08:00 A.M. to 07:59 P.M.). As a dummy variable, off-peak and on-peak are denoted by 0 and 1, respectively. In the regression equation, the coefficient for the dummy variable would indicate how the on-peak attribute has an effect on the dependent variable in reference to the off-peak attribute. The category which is designated as 0 (i.e., off-peak) in the categorical variable is known as the *reference group*.

Consider a NAN consisting of $N$ consumers. In the NAN, each energy thief commits energy theft independently and randomly. Let $\mathbf{x}$ denotes the categorical variables in the model. The period of energy theft or metering defect (i.e., off-peak and on-peak) can be identified by defining another metric known as *detection coefficient*, $\beta$ to the regression equation as follows:

$$\begin{cases} a_1 p_{t_1,1} + \ldots + a_N p_{t_1,N} + \beta_1 p_{t_1,1} x_1 + \ldots + \beta_N p_{t_1,N} x_N = y_{t_1} \\ \vdots \\ a_1 p_{t_T,1} + \ldots + a_N p_{t_T,N} + \beta_1 p_{t_T,1} x_1 + \ldots + \beta_N p_{t_T,N} x_N = y_{t_T}, \end{cases} \tag{3.11}$$

whereby $\beta_n$ indicates whether consumer $n$ cheats inconsistently in a day for $n = 1, 2, \cdots, N$.

Since the category **'off-peak'** is the reference group, it is designated as 0 in the dummy variable. Thus, a LSE is formed to identify random fraudulent consumers who cheat during off-peak hours as follows:

$$a_1 p_{t_o,1} + \ldots + a_N p_{t_o,N} + \beta_1 p_{t_o,1} \cdot 0 + \ldots + \beta_N p_{t_o,N} \cdot 0 = y_{t_o}, \tag{3.12}$$

whereby $p_{t_o,n}$ denotes the energy consumption reported by consumer $n$ during off-peak hours at time interval $t_o \in \{08{:}00 \text{ P.M.}, 08{:}30 \text{ P.M.}, \cdots, 07{:}30 \text{ A.M.}\}$. Note that the time granularity is 30 minutes. Thus, the following equation holds:

University of Malaya

$$a_1 p_{t_o,1} + \ldots + a_N p_{t_o,N} = y_{t_o}, \tag{3.13}$$

for $\forall\ t_o$.

The LSE can also be expressed in matrix-vector form:

$$\mathbf{P^{off}a} = \mathbf{y^{off}}, \tag{3.14}$$

which is similar to Equation (3.6). In Equation (3.14), **a** represents the vector of anomaly coefficients of consumers during off-peak hours.

On the other hand, the group **'on-peak'** is designated as 1 in the dummy variable. Thus, another LSE is formed to detect consumers who perpetrate theft during on-peak hours or faulty SMs as follows:

$$a_1 p_{t_p,1} + \ldots + a_N p_{t_p,N} + \beta_1 p_{t_p,1} \cdot 1 + \ldots + \beta_N p_{t_p,N} \cdot 1 = y_{t_p}, \tag{3.15}$$

which can also be re-arranged as:

$$(a_1 + \beta_1) p_{t_p,1} + \ldots + (a_N + \beta_N) p_{t_p,N} = y_{t_p}, \tag{3.16}$$

whereby $p_{t_p,n}$ denotes the energy consumption reported by consumer $n$ during on-peak hours at time interval $t_p \in \{08{:}00\ \text{A.M.}, 08{:}30\ \text{A.M.}, \cdots, 07{:}30\ \text{P.M.}\}$.

In matrix form, the LSE for the **'on-peak'** group can be expressed by:

$$\mathbf{P^{peak}}(\mathbf{a} + \beta) = \mathbf{y^{peak}}, \tag{3.17}$$

where (**a**+$\beta$) denotes the anomaly coefficients of consumers during on-peak hours. **a** itself



University of Malaya

**Table 3.2: Description of $a$, $\beta$ and $(a + \beta)$**

| Scenario | $a$ | $\beta$ | $a + \beta$ | Description |
|---|---|---|---|---|
| 1 | $= 0$ | $= 0$ | $= 0$ | Honest |
| 2 | $> 0$ | $= 0$ | $> 0$ | Cheating constantly |
| 3 | $< 0$ | $= 0$ | $< 0$ | Faulty constantly |
| 4 | $= 0$ | $> 0$ | $> 0$ | Cheating during on-peak |
| 5 | $= 0$ | $< 0$ | $< 0$ | Faulty during on-peak |
| 6 | $> 0$ | $-a$ | $= 0$ | Cheating during off-peak |
| 7 | $-\beta$ | $> 0$ | $= 0$ | Faulty during off-peak |

denotes the anomaly coefficients of consumers during off-peak period. The coefficient for categorical variable, known as *detection coefficient* (i.e., $\beta$) would indicate how the on-peak attribute has an impact on the dependent response $\mathbf{y}$.

By applying Equation (3.9), the maximum likelihood estimator of the regression coefficients are thus computed by:

$$\begin{bmatrix} \mathbf{a} \\ \hdashline \beta \end{bmatrix} = ((\mathbf{P}^{\mathbf{aug}})'\mathbf{P}^{\mathbf{aug}})^{-1}(\mathbf{P}^{\mathbf{aug}})'\mathbf{y} \tag{3.18}$$

where,

$$\mathbf{P}^{\mathbf{aug}} = \left[\begin{array}{c:c} \mathbf{P}^{\mathbf{off}} & 0 \\ \hdashline \mathbf{P}^{\mathbf{peak}} & \mathbf{P}^{\mathbf{peak}} \end{array}\right]. \tag{3.19}$$

By investigating the estimated $\mathbf{a}$ and $\beta$, the dishonest consumers can be deduced whether they are committing theft either all the time or only during a particular period in a day. The following seven scenarios describe the operation of Equation (3.14) and Equation (3.17) to identify cheating consumers or faulty SMs that occur constantly or occasionally through dummy coding. The possible scenarios of each consumer (i.e., $n = 1, \cdots, N$) are summarized in Table 3.2.

- **Scenario 1**: Obviously, both $a$ and $\beta$ equal to 0 imply that each consumer is honest in his/her energy reporting.

University of Malaya

- **Scenario 2**: When $a$ is positive while $\beta = 0$, the sum of $a$ and $\beta$ is also positive. $\beta = 0$ indicates that the anomaly coefficient is constant throughout the observed period. Therefore, it can be concluded that the consumer is cheating on his/her energy consumption in both off-peak and on-peak periods (all the time).
- **Scenario 3**: If $a$ is negative and $\beta = 0$, the total of $a$ and $\beta$ is also negative. These combinations imply that the SM in the consumer's premise is out of order all the time.
- **Scenario 4**: $a = 0$ and $\beta$ is positive. The positive sum of $a$ and $\beta$ indicates that the consumer is cheating only during on-peak period. $a = 0$ implies that there are no cheating or device failure during off-peak hours. Positive $\beta$ shows that there is a status change from non-cheating during off-peak to cheating during on-peak.
- **Scenario 5**: Meanwhile, $a = 0$ and $\beta < 0$ show that SM is defective during on-peak (i.e., $a + \beta < 0$).
- **Scenario 6**: $a$ is positive while $\beta = -a$ (negative). The resultant of $a$ and $\beta$ is equal to 0. These combinations imply that the consumer is cheating on his/her energy consumption only during off-peak period. He/she does not steal electricity during on-peak because $a + \beta = 0$.
- **Scenario 7**: $\beta$ is positive and $a = -\beta$. In such a case, $a + \beta = 0$, thereby indicating that the SM is faulty during off-peak and is working fine during on-peak times.

Scenarios 5 and 7 are not realistic, but are included here for completeness of discussion.

### 3.6.2 CVLR-ETDM

The flow chart in Figure 3.5 shows the operations in Categorical Variable-enhanced LR-ETDM (CVLR-ETDM). Categorical variables are incorporated in the regression model as dummy variables prior to the invocation of CVLR-ETDM. In this work, there are two

University of Malaya

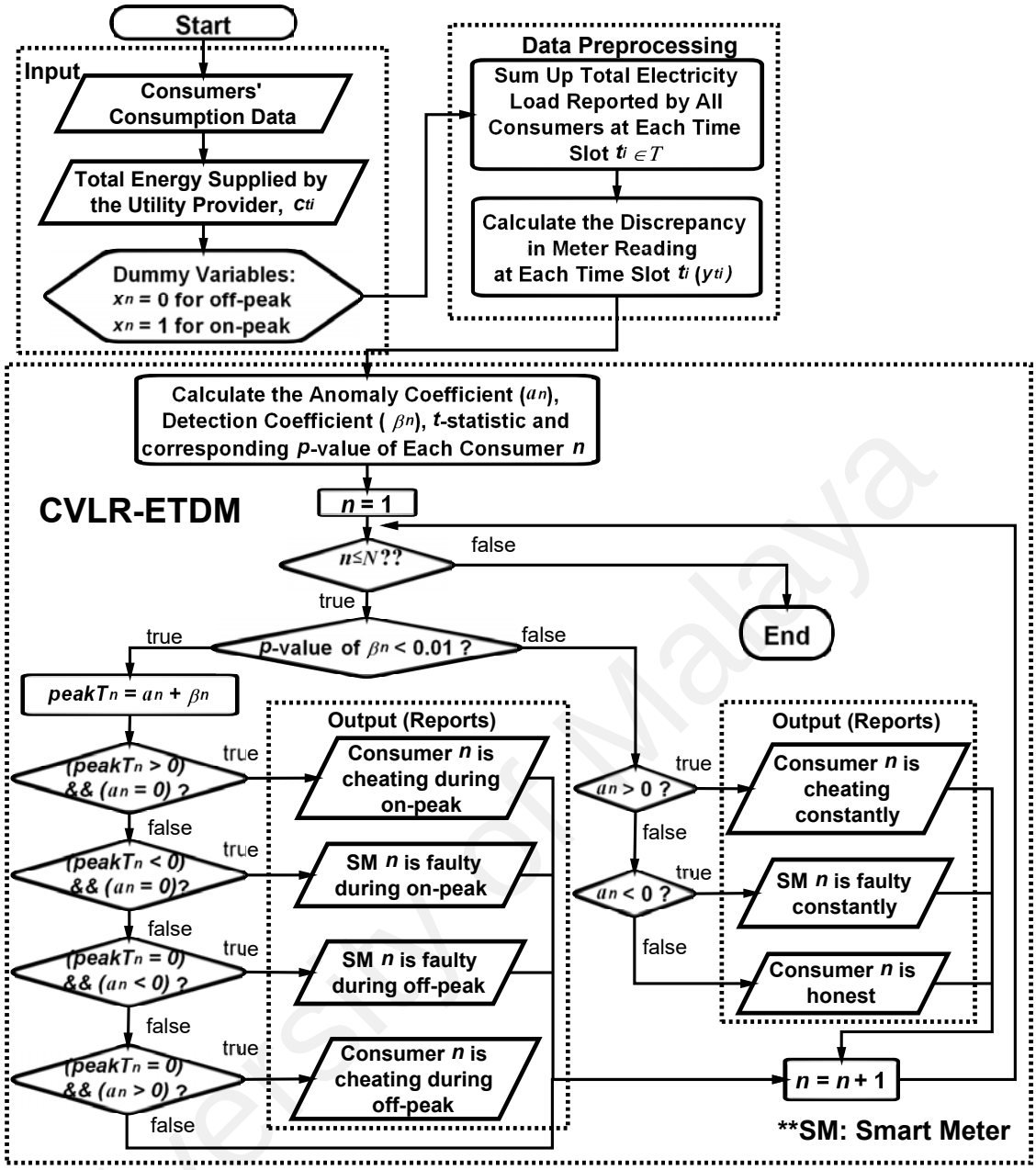



**Figure 3.5: Flow chart of the CVLR-ETDM scheme.**

time attributes (i.e., $k = 2$), namely off-peak and off-peak. Therefore, one dummy variable (i.e., $k - 1 = 1$) is created for each consumer. In total, there are $2N$ coefficients (i.e., $N$ anomaly coefficients and $N$ dummy variables). Recall that, off-peak and on-peak are designated by 0 and 1, respectively.

Next, the $p$-value of $\beta_n$ is verified to test the significance of the coefficient given the other coefficients. If the $p$-value of $\beta_n$ is less than 0.01, it means that the $t$-statistic is significant at the 1% level given the other coefficients. In other words, $\beta_n$ is non-zero

University of Malaya

(i.e., $a_n$ is not constant) and thus consumer $n$ or $n$-th SM has different cheating pattern throughout the period of observation. In such a case, ($peakT_n = a_n + \beta_n$) is computed to solve Equation (3.17) for determining the anomaly coefficient of consumer $n$ during on-peak hours. The outcome of $peakT_n > 0$ and $a_n = 0$ indicates that SM reading of consumer $n$ is reporting less only during on-peak hours. If $peakT_n < 0$ and $a_n = 0$, it implies that the $n$-th SM is malfunctioning during on-peak period. When $peakT_n = 0$ and $a_n < 0$, the $n$-th SM is malfunctioning during off-peak hours. Otherwise, $peakT_n = 0$ and $a_n > 0$ indicate that consumer $n$ steals energy during off-peak period.

On the other hand, the $p$-value of $\beta_n$ greater than 0.01 implies that $a_n$ of consumer $n$ is constant. That is, consumer $n$ cheats or $n$-th SM is faulty consistently throughout the period of observation. In such a case, if $a_n > 0$, it shows that the consumer reports less in his/her energy consumption reporting all the time. Otherwise, the $n$-th SM is out of order when $a_n < 0$. Apart from that, $a_n = 0$ shows that consumer $n$ is honest in reporting his/her electricity consumption.

### 3.6.3 Differences of Data Involved for LR-ETDM and CVLR-ETDM

In this section, graphical illustrations to show the differences of data involved for the computation of LR-ETDM (i.e., Equation (3.5)) and CVLR-ETDM (i.e., Equation (3.11)) are presented to elaborate each MLR-based anomaly detection scheme for the beneficial of reader to understand the proposed LR-based schemes.

Consider a NAN consisting of $N$ consumers. Suppose that the SM readings are sampled over $T$ time interval everyday. Let $p_{t_i,n}$ and $c_{t_i}$ denote the near real-time energy consumption recorded by consumer $n \in \mathbf{N} = \{1, 2, \cdots, N\}$ and collector, respectively, at time interval $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$. Anomaly coefficient, denoted by $a_n$, is defined for each consumer $n$ to assess the reliability of SM endowed in each household or honesty level of each consumer in energy reporting. Recall that $y_{t_i}$ is the discrepancy between the total electricity

University of Malaya

supplied by the UP and the sum of energy consumption reported by all consumers in the service area at time interval $t_i$.

As initial work, the constant scenario is first considered whereby energy thieves never stop cheating and/or defective meters are faulty all the time. Therefore, the rate and pattern of cheating/malfunctioning remain the same throughout the period of observation. The data involved to solve the LSE in Equation (3.5) for the detection of *constant* anomaly coefficients using LR-ETDM is graphically represented in Figure 3.6. One-day half-hourly metered energy consumption data (i.e., 48 data points, highlighted in blue) are extracted for the detection of constant anomaly coefficients as long as the number of observations is greater than the number of consumers in the service area (i.e., $T > N$). Here, the $t_i$-th row of the table in Figure 3.6 represents the data recorded by all $N$ consumers and meter discrepancy at the $t_i$-th time interval. On the other hand, the $n$-th column and the last column of the table in Figure 3.6 denote the data measured by the SM for consumer $n$ and the meter discrepancies, respectively, over all $t_i$. For service area of larger size when the number of consumers is more than the number of observations (i.e., $N > T$), the energy consumption data are observed over longer period to improve the detection rate.

However, some of the energy thieves and/or defective meters evade detection when NTLs occur only during a certain period in a day, as discussed in Section 3.6. Specifically, LR-ETDM tends to produce an incorrect anomaly coefficient vector where some of the predicted elements show inaccurate values when dishonest consumers attempt to falsify their energy consumption inconsistently. To overcome the shortcoming of the LR-ETDM scheme, a Categorical Variable-Enhanced LR-ETDM (CVLR-ETDM) is put forward to solve varying anomaly coefficients. The data involved to solve the LSE in Equation (3.11) for the detection of *varying* anomaly coefficients using CVLR-ETDM is illustrated in Figure 3.6. In this work, categorical variable of consumer $n$ ($x_n$) is introduced to the

University of Malaya

**LR-ETDM**

| Time Interval | Consumer 1 | Consumer 2 | ... | Consumer *N* | Discrepancy in Meter Reading |
|---|---|---|---|---|---|
| 1st interval, $t_1$ | $a_1 p_{t_1,1}$ | $a_2 p_{t_1,2}$ | ... | $a_N p_{t_1,N}$ | $y_{t_1}$ |
| | | | ⋮ | | |
| 15th interval, $t_{15}$ | $a_1 p_{t_{15},1}$ | $a_2 p_{t_{15},2}$ | ... | $a_N p_{t_{15},N}$ | $y_{t_{15}}$ |
| 16th interval, $t_{16}$ | $a_1 p_{t_{16},1}$ | $a_2 p_{t_{16},2}$ | ... | $a_N p_{t_{16},N}$ | $y_{t_{16}}$ |
| | | | ⋮ | | |
| 39th interval, $t_{39}$ | $a_1 p_{t_{39},1}$ | $a_2 p_{t_{39},2}$ | ... | $a_N p_{t_{39},N}$ | $y_{t_{39}}$ |
| 40th interval, $t_{40}$ | $a_1 p_{t_{40},1}$ | $a_2 p_{t_{40},2}$ | ... | $a_N p_{t_{40},N}$ | $y_{t_{40}}$ |
| | | | ⋮ | | |
| 48th interval, $t_{48}$ | $a_1 p_{t_{48},1}$ | $a_2 p_{t_{48},2}$ | ... | $a_N p_{t_{48},N}$ | $y_{t_{48}}$ |

Anomaly Detection for *Constant* Anomaly Coefficients

**Figure 3.6: Graphical illustration to show the data involved for the computation of the LR-ETDM scheme.**

framework to categorize the time of under-reporting/over-reporting by *n*-th SM into two groups, namely off-peak (i.e., from 1st to 15th time intervals and from 40th to 48th time intervals, highlighted in red) and on-peak (i.e., from 16th to 39th time intervals, highlighted in orange). As shown in Figure 3.7, the categorical variables for off-peak and on-peak are denoted by 0 and 1, respectively. Besides that, another metric known as detection coefficient, $\beta_n$ is also defined to determine whether consumer *n* cheats on energy reporting or the *n*-th SM is out of order inconsistently in a day. The change in cheating/malfunctioning behavior may be quickly deduced from the detection coefficient of consumer *n* ($\beta_n$). Similar to LR-ETDM, one-day metered energy consumption data are required for the detection analysis as long as the number of observations is more than the number of consumers in the NAN. For service area of larger size, SM readings are observed over longer period to increase the number of observations so as to mitigate the effect of over-fitting (Tetko, Livingstone, & Luik, 1995).

University of Malaya

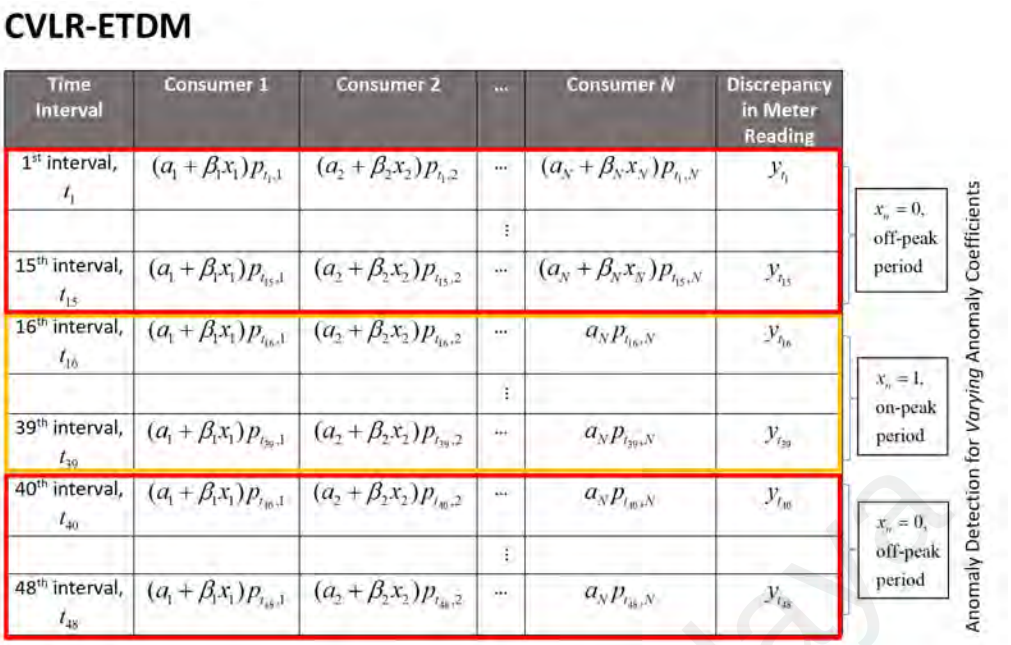



**Figure 3.7: Graphical illustration to show the data involved for the computation of the CVLR-ETDM scheme.**

## 3.7 Summary of Chapter

This chapter puts forward two novel detection schemes, namely LR-ETDM and CVLR-ETDM, to study consumers' energy utilization behavior and evaluate their honestly level in energy reporting, with the aim to reduce revenue losses and costs incurred due to NTLs. The proposed schemes are based on MLR. Any non-zero anomaly coefficients are indicative of energy thefts or metering defects. It is observed that LR-ETDM might be unstable when there are inconsistent energy thefts and/or defective SMs. Therefore, categorical variables are incorporated into MLR and CVLR-ETDM is developed so that the framework can successfully detect consumers' malfeasance and faulty SMs even when there are inconsistent cheating/malfunctioning events.

To enhance detection accuracy and minimize false positives, more attention will be devoted to handle the noise tolerance issue of the proposed schemes. Specifically, the impact caused by TLs and measurement noise/error on the detection analysis will be taken into consideration in the design of NTL detection scheme in the next chapter.

University of Malaya

# CHAPTER 4: LINEAR PROGRAMMING-BASED ANOMALY DETECTION FRAMEWORK

## 4.1 Overview

This chapter provides the methodology proposed for the LP-based anomaly detection framework and implements the associated key algorithms to be used for improving detection accuracy and reducing false positives. The first sub-chapter presents the rationale behind the new proposed anomaly detection framework. The next sub-chapter investigates the impact caused by TLs and measurement noise/error on the anomaly detection analysis. Subsequently, an optimization framework which takes into account the impact caused by TLs and measurement noise on the design of anomaly detection framework, which includes the energy balance analysis and fraction of reported consumption computation is presented. Next, the problem formulation is discussed, and the flow processes of the proposed LP-based detection schemes are elaborated in the following sub-chapters. Lastly, a summary is given to conclude the proposed anomaly detection schemes.

## 4.2 Motivation

Although there exist some intelligent schemes to study consumers' energy consumption behavior for NTL detection in smart grids such as (S. Salinas et al., 2013; Jokar et al., 2016), their proposals did not consider the impact caused by TLs and measurement noise on the energy theft detection. According to Sahoo et al. (2015), an effective way of detecting NTLs in the distribution network is by correctly estimating TLs in the network. In the previous work as discussed in Chapter 3, two energy theft detection schemes utilizing MLR are proposed. However, these schemes assume that power line losses are known, which in practice may be difficult to acquire. In the real world, calibration error and TLs should be considered in the design of the anomaly detection frameworks. In error analysis, the energy balance error is the sum of the errors from both DS and all the consumers' meters. The

University of Malaya

consumers' meter errors are usually small and tend to cancel each other out. However, the meter errors of the DS, are larger and influence the overall energy balance error. According to Accenture (2011), for DS consisting of 50 consumers, the energy balance error is usually less than 240 watts (W) when there is no downstream energy. Any energy balance signal greater than 240 W is therefore almost certainly not measurement error but legitimate missing energy. In the pursuit of higher anomaly detection rate, the assumption of known power line losses is relaxed and a new LP-based anomaly detection framework that can overcome the deficiency of the previous work is proposed in this chapter. Particularly, the impact caused by TLs and measurement noise/error on the detection analysis is taken into account to enhance the detection rate and minimize false positives.

## 4.3 Impact of TLs and Measurement Noise/Error on NTL Detection Analysis

Generally, a substantial amount of energy is lost in the power grid due to both TLs and NTLs. NTLs primarily relate to energy theft and meter irregularities. As discussed in Section 2.3.2, energy theft includes a number of methods to deliberately defraud the UPs. On the other hand, TLs comprise ohmic losses in power grid caused by the line resistances, leaking due to imperfect isolation, conversion losses at the DT, etc (Xu, 2015). The total of TLs varies significantly throughout the day, week, month and year as some of the components of TLs are subject to the amount of power being delivered to the consumers (Nikovski et al., 2013). Therefore, it is challenging to determine whether the losses is either technical or non-technical (i.e., theft).

Ideally, the electrical energy generated by the UPs should be equal to the total energy reported by all the consumers in a service area. However, in practice, these two amounts will not tally because losses occur as an integral result of energy transmission and distribution. The actual losses are the difference between outgoing energy recorded by the collector at the DS and total energy billed to the consumers. The discrepancy between actual losses

University of Malaya

and expected losses would yield the extent of NTLs in the service area.

It might be possible to estimate the percentage of TLs accurately using load flow analysis if all the parameters of the distribution network were known, including the order and attachment points of all consumers, the line resistances between the attachment points, connection topology and the instantaneous power consumption of every consumer (Navani, Sharma, & Sapra, 2012). Nonetheless, in practice, full knowledge of these parameters is not feasible. For instance, UPs only know the DT serving each consumer but not the exact line resistances or the connection order. Besides that, the instantaneous power consumed by each consumer at any time instant can be determined only by installing detailed measurement device, which records fine-grained measurements such as phasor measurement units (PMUs) at each consumer's premise. However, installation of such device would impose higher cost to the UPs, even far exceeding the cost incurred due to energy theft.

Before the introduction of SG, UPs collect aggregated measurements, usually over one-month with the conventional analogue meters. With the advancement of SM, UPs are able to predict the percentage of TLs more accurately and detect NTLs with shorter duration based on the near real-time measurements. Nonetheless, the fine-grained scale of analysis will not improve the detection rate significantly when the presence of TLs and measurement noise is ignored. Specifically, if the impact caused by TLs and measurement noise on the analysis is ignored, higher accuracy of detection might not be expected for the systematic long-term energy theft. In other words, the proposed anomaly detection framework might wrongly accuses some of the honest consumers as fraudulent ones, results in a significant number of false positives in detection. Thus, it is crucial to distinguish TL from the NTL components to accurately detect energy thefts. In order to improve the detection accuracy and reduce false positives, TLs as well as measurement noise/error

University of Malaya

must be taken into consideration and more advanced intelligent detection schemes are needed in the design of anomaly detection framework.

## 4.4 LP Model for Detecting NTLs

In this section, the mathematical models for detecting energy frauds and defective SMs in a NAN are presented.

### 4.4.1 Energy Balance Analysis

One of the theft detection feature of the state-of-the-art MDMS is the ability to estimate TLs. TLs can be estimated by performing energy balance between the total power supplied by the DS and the total energy consumed by all metered consumers. In MDMS, the total of TLs, which is also known as the loss rate, is calculated as a percentage of the total supplied power (Nikovski et al., 2013). If the loss rate exceeds a certain threshold (e.g., 3%), energy theft is suspected and alarm will be triggered.

The problem of detecting NTLs in a branch of radial distribution network which consists of a DT connected to a DS and a number of consumers connected to the secondary side of the DT as shown in Figure 3.2 is studied. Consider a service area consisting of $N$ consumers. Let $p_{t_i,n}$ and $c_{t_i}$ denote the near real-time energy consumption recorded by consumer $n$ and data collector, respectively, at time interval $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$. Similar to the previous work in Chapter 3, an anomaly coefficient, denoted by $a_n$, is defined for each consumer such that $a_n = 0$ if consumer $n$ is truthful in reporting his energy consumption. Any non-zero $a_n$ is indicative of energy fraud or faulty SM. As discussed in Chapter 3, since most power distribution networks are radial in practice, only radial topology as shown in Figure 3.3 is considered. Recall that, the leaf nodes (i.e., consumers) are connected to the root node (i.e., DS). Besides, the losses (i.e., both TLs and NTLs) are also modeled as leaf nodes. As active power is additive, the total energy supplied by

University of Malaya

the root node to the NAN should equal the sum of electricity consumption reported by all the leaf nodes at time interval $t_i$ (Rashed Mohassel et al., 2014). Therefore, the following equation is formulated:

$$\sum_{n=1}^{N} p_{t_i,n} + \lambda + \theta + \gamma = c_{t_i}, \tag{4.1}$$

where $\lambda$ denotes the TLs, while $\theta$ and $\gamma$ indicate inaccurate meter readings due to energy thefts and faulty SMs, respectively.

Leveraging the loss rate computed by the MDMS, another metric known as the loss factor, $l_{t_i}$ is introduced in Equation (4.1) to capture the amount of TLs, denoted by $\lambda$, during each time interval. Here, $\lambda$ is assumed to be proportional to $c_{t_i}$ with the proportion coefficient $l_{t_i}$. Specifically, the following equation holds:

$$\lambda = l_{t_i} c_{t_i}. \tag{4.2}$$

When TLs are taken into account, high accuracy of NTL detection due to electricity pilfering and faulty SMs could be achieved even with the existence of sophisticated energy theft.

Therefore, Equation (4.1) can also be mathematically formulated as:

$$(a_1 + 1)p_{t_i,1} + (a_2 + 1)p_{t_i,2} + \ldots + (a_N + 1)p_{t_i,N} + l_{t_i} c_{t_i} = c_{t_i}. \tag{4.3}$$

Re-arranging which gives:

$$a_1 p_{t_i,1} + a_2 p_{t_i,2} + \ldots + a_N p_{t_i,N} + l_{t_i} c_{t_i} = c_{t_i} - \sum_{n=1}^{N} p_{t_i,n}. \tag{4.4}$$

As mentioned earlier, in the event there is neither energy fraud nor faulty meter, the total

University of Malaya

power supplied by UPs should be approximately the same as the sum of metered consumers' energy consumption measured for every time interval in a NAN. Slight differences will be due to TLs (e.g., un-metered loads, line losses, etc.) connected to DS, but these are relatively small in comparison to the NTLs caused by energy diversion.

Therefore, with reference to Equation (4.1), for the scenario where $\theta > 0$ (i.e., energy theft exists) as well as $\gamma < 0$ (i.e., at least one SM is faulty), the discrepancy in meter reading at time $t_i$, denoted by $y_{t_i}$, is computed as:

$$y_{t_i} = c_{t_i} - \sum_{n=1}^{N} p_{t_i,n} = \lambda + \theta + \gamma, \tag{4.5}$$

where $\lambda = l_{t_i} c_{t_i} > 0$ shows that TLs are present in the NAN, $\theta = \sum_{\substack{n=1 \\ a_n>0}}^{N} a_n p_{t_i,n}$ indicates energy theft and $\gamma = \sum_{\substack{n=1 \\ a_n<0}}^{N} a_n p_{t_i,n}$ represents defective meters.

Suppose that the electricity consumption is sampled over $T$ time intervals within a day. A LSE for the detection of electricity theft and defective meters can be formulated as follows:

$$\begin{cases} a_1 p_{t_1,1} + a_2 p_{t_1,2} + \ldots + a_N p_{t_1,N} + l_{t_1} c_{t_1} &= y_{t_1} \\ a_1 p_{t_2,1} + a_2 p_{t_2,2} + \ldots + a_N p_{t_2,N} + l_{t_2} c_{t_2} &= y_{t_2} \\ & \vdots \\ a_1 p_{t_T,1} + a_2 p_{t_T,2} + \ldots + a_N p_{t_T,N} + l_{t_T} c_{t_T} &= y_{t_T} \end{cases} \tag{4.6}$$

Similar to previous work, the goal is to find all $a_n, n \in \mathbf{N} = \{1, 2, \cdots, N\}$ of the LSE in Equation (4.6) for evaluating the anomalous behavior of each consumer or reliability of SM endowed in each household. Indeed, there are three possibilities as described in Table 3.1 of Chapter 3.

University of Malaya

### 4.4.2 Fraction of Reported Consumption

Besides detecting anomalies in the consumption patterns, the operation center is also able to compute the fraction of reported consumption (i.e., amount of under-reporting/over-reporting) of each consumer based on the computed anomaly coefficient. Recall that, the sum of electricity consumption reported by all the consumers, $\sum_{n=1}^{N} p_{t_i,n}$ must agree with the total load consumption measured by the collector at time interval, $c_{t_i}$ as discussed in Equation (4.3). The computation for fraction of reported consumption is explained using a one-consumer topology, namely the $n$-th consumer. For simplicity, it is assumed that there is neither noise in the measurement nor TLs (i.e., $l_{t_i} = 0$). Equation (4.3) becomes $(a_n + 1)p_{t_i,n} = c_{t_i}$. Therefore, the fraction of reported consumption of the $n$-th consumer is computed as follows:

$$\frac{1}{a_n + 1} = \frac{p_{t_i,n}}{c_{t_i}} \tag{4.7}$$

When the consumer is honest in energy reporting, the reported energy consumption of the consumer will tally with the total energy supplied by the collector (i.e., $p_{t_i,n} = c_{t_i}$). Thus, $\frac{1}{a_n+1} = 1$. On the other hand, $\frac{1}{a_n+1} < 1$ when the consumer under-reports what was consumed (i.e., $p_{t_i,n} < c_{t_i}$). Similarly, $\frac{1}{a_n+1} > 1$ when the consumer over-reports his/her energy consumption (i.e., $p_{t_i,n} > c_{t_i}$).

Meanwhile, to compute the fraction of reported consumption of each consumer either during off-peak period or on-peak hours using CVLR-ETDM, the detection coefficient $\beta_n$ in Equation (4.3), as discussed in Chapter 3 is included. Without the presence of measurement noise and TLs, Equation (4.3) becomes $(a_n + \beta_n + 1)p_{t_i,n} = c_{t_i}$. Recall that, $a_n$ itself denotes the anomaly coefficient of consumer $n$ during off-peak period while $(a_n + \beta_n)$ denotes the anomaly coefficient of consumer $n$ during on-peak hours. Therefore, the fraction of reported energy usage for each consumer during off-peak period is calculated in

University of Malaya

a manner similar to Equation (4.7). On the other hand, the fraction of reported consumption of consumer $n$ during on-peak hours is computed as follows:

$$\frac{1}{(a_n + \beta_n) + 1} = \frac{p_{t_i,n}}{c_{t_i}} \tag{4.8}$$

Based on the values of $(\frac{1}{a_n+1} \times 100\%)$ or $(\frac{1}{(a_n+\beta_n)+1} \times 100\%)$, the operation center is able to estimate the percentage of under-reporting or over-reporting by each SM. Therefore, the operation center can easily detect fraudulent consumers and discover defective SMs in the NAN by referring to the computed anomaly coefficient, detection coefficient and fraction of reported consumption.

## 4.5 Problem Formulation

In this subsection, an **A**nomaly **D**etection **F**ramework, hereinafter abbreviated as **ADF**, is developed to identify potential energy thieves and faulty SMs in a NAN. In Sections 4.5.2 and 4.5.3, two schemes are presented to solve the LSE for the anomaly coefficients using LP. LP is chosen instead of MLR as discussed in the previous work in Chapter 3 because of the non-multicollinearity characteristic of MLR. In other words, MLR cannot be used to solve the LSE in Equation (4.6) when the predictors are significantly correlated due to the fact that $c_{t_i} \approx p_{t_i,1} + p_{t_i,2} \cdots + p_{t_i,N}$. When multicollinearity is present, MLR is unable to estimate the coefficients accurately (Studenmund, 2014).

### 4.5.1 Linear Programming

LP, involves minimizing or maximizing a linear objective function subject to bounds, linear equality, and inequality constraints (Boyd & Vandenberghe, 2004). It is useful in energy, operations research, finance and other areas where relationships between variables can be expressed linearly. Generally, LP is the mathematical problem of finding a vector **x** that minimizes the function:



University of Malaya

$$
\begin{aligned}
\text{minimize} \quad & \mathbf{c}^T\mathbf{x} \\
\text{subject to} \quad & A_{ineq}\mathbf{x} \leq \mathbf{b_{ineq}}, \\
& A_{eq}\mathbf{x} = \mathbf{b_{eq}}, \\
& \mathbf{lb} < x < \mathbf{ub}.
\end{aligned}
\tag{4.9}
$$

Here, $A_{ineq} \in \mathbf{R}^{mxn}$, $\mathbf{b_{ineq}} \in \mathbf{R}^m$, $A_{eq} \in \mathbf{R}^{pxn}$, $\mathbf{b_{eq}} \in \mathbf{R}^p$ and $\mathbf{c} \in \mathbf{R}^n$. The term $\mathbf{c}^T\mathbf{x}$ is known as the objective function. The inequality $A_{ineq}\mathbf{x} \leq \mathbf{b_{ineq}}$ and equality $A_{eq}\mathbf{x} = \mathbf{b_{eq}}$ are the linear constraints while **lb** and **ub** are the lower and upper bound constraints. Linear programs are convex optimization problems, whereby a linear program can be solved efficiently by different methods such as *simplex methods*, *interior point methods* and *active-set methods* (Xu, 2015). Since an affine objective $\mathbf{c}^T\mathbf{x}$ can be maximized by minimizing $-\mathbf{c}^T\mathbf{x}$, a maximization problem with affine objective and constraint functions is also referred to as a linear program. Generally, there is no closed form solution to Equation (4.9). The feasible set of the LP problem in Equation (4.9) is a polyhedron, $P$ as illustrated in Figure 4.1. The objective $\mathbf{c}^T\mathbf{x}$ is linear. Therefore, its level curves are hyper-planes orthogonal to $\mathbf{c}$ (i.e., shown as dashed lines) in the figure. The point $\mathbf{x}^*$ is optimal. It is the furthest point in $P$ in the direction $-\mathbf{c}$.

The following methods are widely used to solve LP problems (Boyd & Vandenberghe, 2004; Koberstein, 2008):

1. **Dual-Simplex**: The dual-simplex method is the most widely used algorithms for LP in Matlab. It utilizes a systematic procedure to generate and test candidate vertex solutions to a linear program.
2. **Interior point**: The interior point method is very useful for large-scale linear programs that can be defined using sparse matrices or have structure. It utilizes a primal-dual predictor-corrector algorithm, and;

University of Malaya

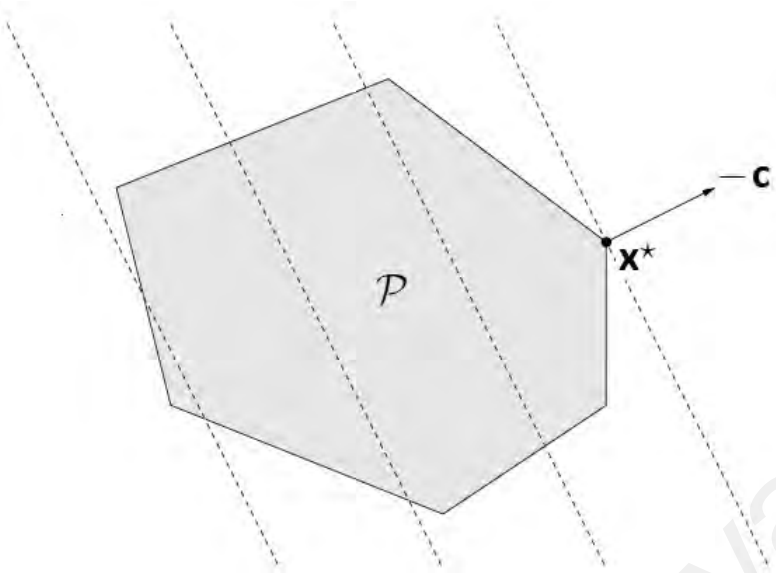



**Figure 4.1: Geometric interpretation of a LP. The shaded region, which is a polyhedron, is the feasible set $P$ (Boyd & Vandenberghe, 2004).**

3. **Interior point-legacy**: The interior point-legacy method is similar to interior point method. However, interior point-legacy method is less robust, slower and consume more memory.

### 4.5.2 Solving Constant Anomaly Coefficients using ADF

Here, a constant scenario where the dishonest consumers always steal energy throughout the entire day is first assumed. Meanwhile, faulty SMs are assumed to always report more than what the corresponding consumers actually consumed. That is, the rate and pattern of cheating/malfunctioning remain the same throughout the period of observation. A LP problem in minimizing the errors of LSE in Equation (4.6) is formulated to solve the anomaly coefficients and loss factors. For this purpose, a new metric known as the *error term*, denoted by $E_{t_i}$ is introduced to each equation in Equation (4.6) for $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$ to capture the random calibration error/measurement noise of the equipment. An accurate and efficient detection of energy theft/metering irregularities can be formulated in terms of minimizing the measurement error/noise at each time interval, which can be expressed as the following optimization problem:

University of Malaya

$$\begin{aligned} \text{minimize} \quad & f = \sum_{i=1}^{T} |E_{t_i}| \\ \text{subject to} \quad & \sum_{n=1}^{N} a_n p_{t_i,n} + l_{t_i} c_{t_i} + E_{t_i} = y_{t_i}, \forall t_i \in \mathbf{T}, \\ & E_{t_i}, a_n \text{ unrestricted}, \forall n \in \mathbf{N}, \forall t_i \in \mathbf{T}. \end{aligned} \tag{4.10}$$

The variable $E_{t_i}$ is further expressed as the difference between two non-negative variables, namely, $(E^+)_{t_i}$ and $(E^-)_{t_i}$, and let $E_{t_i} = (E^+)_{t_i} - (E^-)_{t_i}$. Specifically, $E_{t_i}$ is split into $(E^+)_{t_i}$ and $(E^-)_{t_i}$ to capture the positive and negative calibration/measurement error, respectively. Thus, for $E_{t_i} < 0$ at the optimal stage, then $(E^+)_{t_i} = 0$ and $(E^-)_{t_i} > 0$. On the other hand, when $E_{t_i} > 0$ at the optimal stage, then $(E^+)_{t_i} > 0$ and $(E^-)_{t_i} = 0$. With $E_{t_i}$ replaced by $(E^+)_{t_i} - (E^-)_{t_i}$, the given LP problem is equivalent to the following:

$$\begin{aligned} \text{minimize} \quad & f = \sum_{i=1}^{T} \left( (E^+)_{t_i} + (E^-)_{t_i} \right) \\ \text{subject to} \quad & \sum_{n=1}^{N} a_n p_{t_i,n} + l_{t_i} c_{t_i} + (E^+)_{t_i} - (E^-)_{t_i} = y_{t_i}, \forall t_i \in \mathbf{T}, \end{aligned} \tag{4.11}$$

$$(E^+)_{t_i}, (E^-)_{t_i} \geq 0, \forall t_i \in \mathbf{T}, \tag{4.12}$$

$$a_n \text{ unrestricted}, \forall n \in \mathbf{N}, \tag{4.13}$$

$$l_{min} \leq l_{t_i} \leq l_{max}, \forall t_i \in \mathbf{T}. \tag{4.14}$$

In general, a lower $f$ is preferred for higher accuracy in locating the potential energy fraudsters and/or defective SMs. As discussed in Section 4.5.1, it can be solved in a centralized fashion by using either the dual-simplex method or the interior point method, which is packaged in the Optimization Toolbox of Matlab R2014b. Note that the design

University of Malaya

of the TL estimator is not the focus of this thesis. However, the range of TLs can be estimated based on measurements at the collector and the knowledge of the distribution network (Au et al., 2008; Sahoo et al., 2015; Oliveira & Padilha-Feltrin, 2009). For LV network, TLs are primarily influenced by its load factor, loading, capacity and network type (i.e., overhead or underground). According to Au et al. (2008), the average TLs of LV network in Malaysia was reported to range from 0.59% to 3.23%. To show the viability of the proposed framework in the presence of TLs, it is assumed that there are $3\% - 5\%$ TLs in a NAN (i.e., $l_{min} = 0.03$, $l_{max} = 0.05$) as captured in the constraint expressed as Equation (4.14). The values of $l_{min}$ and $l_{max}$ are configured based on the estimated range of TLs in the NAN. The flow chart as shown in Figure 4.2 summarizes the Anomaly Detection Framework (ADF) scheme.

Assume that the collector labels the SM of all consumers in the NAN of interest from 1 to $N$. Before ADF is invoked, data cleaning is performed to filter out suspicious SMs which do not have measurements, report a constant low value all the time or have corrupted data. Correspondingly, the faulty or compromised SMs should be inspected and replaced so that ADF can obtain a more accurate NTL detection analysis. The scheme commences by computing the discrepancies between the total metered energy consumption of all consumers in the service area and the total power supplied by the UPs for time interval $t_i \in \mathbf{T}$ as shown in Equation (4.5). Then, a LSE consisting of consumers' reported meter data, anomaly coefficients, loss factors, total supplied power and the differences in reading is formed as expressed by Equation (4.6). The `linprog` function packaged in the Optimization Toolbox of Matlab R2014b is deployed to solve the LP for anomaly coefficients $a_n$, loss factors $l_{t_i}$ as well as error terms $(E^+)_{t_i}$ and $(E^-)_{t_i}$.

In this subsection, it is assumed that either the fraudulent consumer *always* under-reports his/her energy reporting or the faulty SM over-reports what was consumed by more than 5%,

University of Malaya

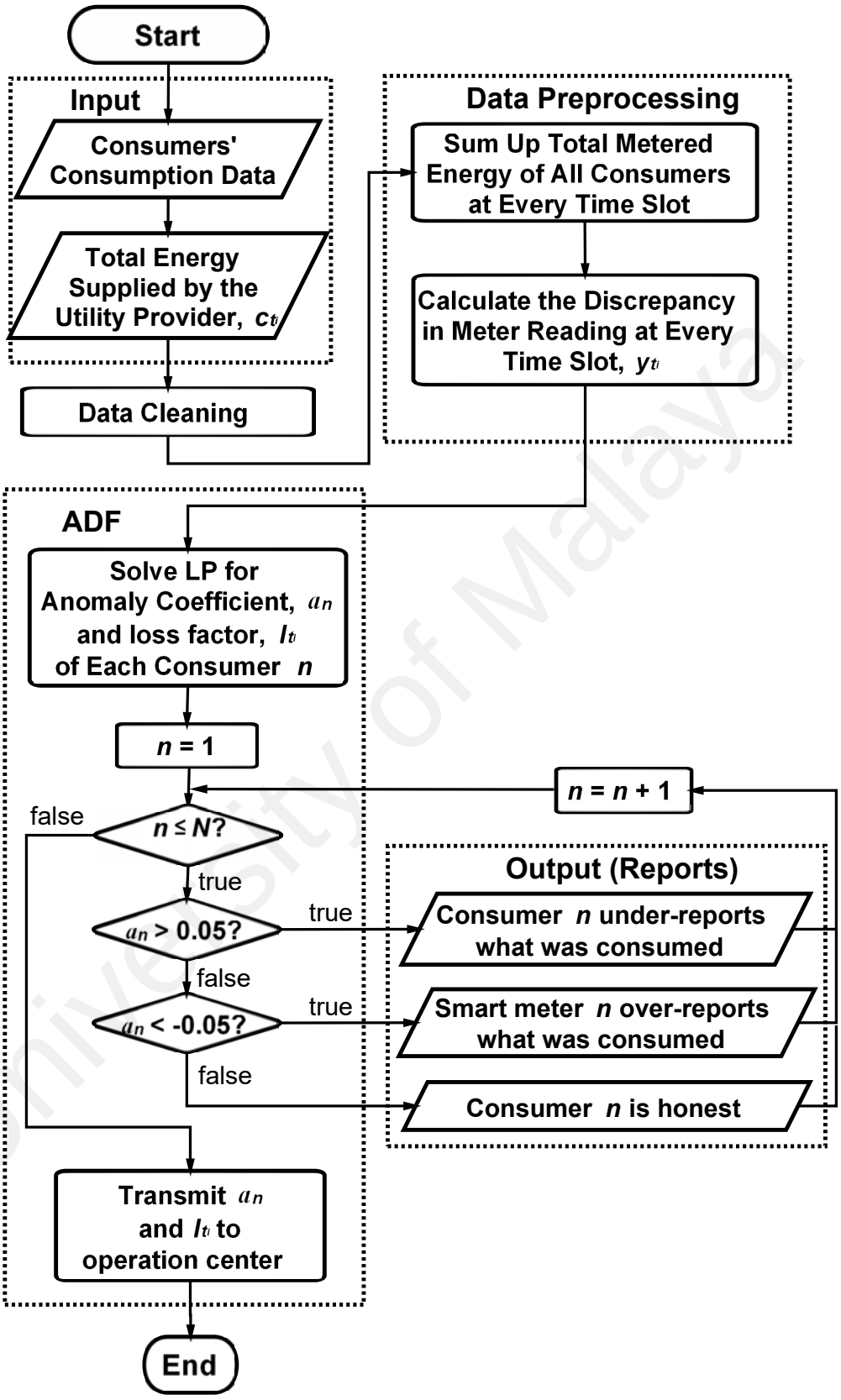



**Figure 4.2: Flow chart of the ADF scheme.**

University of Malaya

respectively. In other words, consumers who have anomaly coefficients in $[-0.05, 0.05]$ are assumed to be honest (i.e., $a_n \approx 0$). The slight differences between the exact and computed anomaly coefficients are likely due to the calibration error/measurement noise. The reliability of SM or anomalous behavior of each consumer is evaluated by solving $a_n$ for $n \in \mathbf{N}$ as discussed in Table 3.1. Subsequently, the collector will transmit the computed consumers' anomaly coefficient, $a_n$ and percentage of TLs, $l_{t_i}, \forall t_i \in \mathbf{T}$ to the operation center. Note that the collector invokes ADF at the end of each day. For service area of larger size, consumers' power consumption data are observed over longer period to increase the detection accuracy.

### 4.5.3 Solving Varying Anomaly Coefficients using Enhanced ADF

As preliminary work in Section 4.5.2, the anomaly coefficients $a_1, a_2, \cdots, a_N$ are assumed to be constant under the ADF scheme. In other words, energy thieves steal energy at the same rate and never stop cheating throughout the period of observation. However, it is possible that the stealing rate varies over time (S. Salinas et al., 2013). Specifically, a fraudulent consumer can manipulate the SM in such a way that they steal energy at different rates during random periods. Consequently, some of the fraudulent consumers/faulty meters stay undetected by ADF when electricity pilfering/meter irregularities occur only during a certain period in a day.

To overcome the deficiency of the ADF scheme, the assumption of constant anomaly coefficients is removed and an Enhanced ADF scheme is put forward to solve more diverse and sophisticated attack types as discussed in the work by Jokar et al. (2016). In such a case, the consumers' reported consumption readings are analyzed over a longer period (i.e., at least $N$ days, where $N$ is the size of the service area) to identify the locations and periods of the consumers' malfeasance and meter irregularities when NTLs take place *all the time* and/or *at varying rates during random intermittent periods in a day*.

University of Malaya

Consider a service area consisting of $N$ consumers. Suppose that the consumers' SM readings are sampled over $T$ time intervals everyday for a period of $D$ days. Let $p^d_{t_i,n}$ and $a_{t_i,n}$ denote the energy consumption recorded by consumer $n$ on day $d \in \mathbf{D} = \{1, 2, \cdots, D\}$ and the anomaly coefficient of consumer $n$, respectively, at time interval $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$. Meanwhile, let $c^d_{t_i}$ and $l^d_{t_i}$ denote the total energy supplied by the UPs and the loss factor at time interval $t_i$, respectively, on day $d$. Therefore, the meter discrepancy at time interval $t_i$ on day $d$ ($y^d_{t_i}$) is computed as:

$$y^d_{t_i} = c^d_{t_i} - \sum_{n=1}^{N} p^d_{t_i,n}. \tag{4.15}$$

Then, a LSE for the detection of varying anomaly coefficients can be formed as follows:

$$\begin{cases} a_{t_1,1}p^1_{t_1,1} + a_{t_1,2}p^1_{t_1,2} + \ldots + a_{t_1,N}p^1_{t_1,N} + l^1_{t_1}c^1_{t_1} = y^1_{t_1} \\ a_{t_2,1}p^1_{t_2,1} + a_{t_2,2}p^1_{t_2,2} + \ldots + a_{t_2,N}p^1_{t_2,N} + l^1_{t_2}c^1_{t_2} = y^1_{t_2} \\ \vdots \\ a_{t_T,1}p^1_{t_T,1} + a_{t_T,2}p^1_{t_T,2} + \ldots + a_{t_T,N}p^1_{t_T,N} + l^1_{t_T}c^1_{t_T} = y^1_{t_T} \\ a_{t_1,1}p^2_{t_1,1} + a_{t_1,2}p^2_{t_1,2} + \ldots + a_{t_1,N}p^2_{t_1,N} + l^2_{t_1}c^2_{t_1} = y^2_{t_1} \\ a_{t_2,1}p^2_{t_2,1} + a_{t_2,2}p^2_{t_2,2} + \ldots + a_{t_2,N}p^2_{t_2,N} + l^2_{t_2}c^2_{t_2} = y^2_{t_2} \\ \vdots \\ a_{t_T,1}p^D_{t_T,1} + a_{t_T,2}p^D_{t_T,2} + \ldots + a_{t_T,N}p^D_{t_T,N} + l^D_{t_T}c^D_{t_T} = y^D_{t_T} \end{cases} \tag{4.16}$$

To detect the energy thieves who steal energy at varying rates, the consumers' reported SM readings will be analyzed over a longer period according to specific time slot $t_i$ until the computed values converge. In other words, the reported SM readings are extracted according to time interval $t_i \in \mathbf{T}$ of each day. Consider a service area consisting of $N$ consumers and the consumers' metered data are observed over a week (i.e., $D = 7$ days).

University of Malaya

The LSE is formulated according to time slot $t_i$ of each day as follows:

$$\begin{cases} a_{t_i,1}p^1_{t_i,1} + a_{t_i,2}p^1_{t_i,2} + \cdots + a_{t_i,N}p^1_{t_i,N} + l^1_{t_i}c^1_{t_i} &= y^1_{t_i} \\ a_{t_i,1}p^2_{t_i,1} + a_{t_i,2}p^2_{t_i,2} + \cdots + a_{t_i,N}p^2_{t_i,N} + l^2_{t_i}c^2_{t_i} &= y^2_{t_i} \\ & \vdots \\ a_{t_i,1}p^7_{t_i,1} + a_{t_i,2}p^7_{t_i,2} + \cdots + a_{t_i,N}p^7_{t_i,N} + l^7_{t_i}c^7_{t_i} &= y^7_{t_i}. \end{cases} \tag{4.17}$$

Thus, the varying anomaly coefficients can be determined more accurately by solving the following objective function:

For each $t_i \in \mathbf{T}$,

$$\begin{aligned} \text{minimize} \quad & f = \sum_{d=1}^{D} \left( (E^+)^d_{t_i} + (E^-)^d_{t_i} \right) \\ \text{subject to} \quad & \sum_{n=1}^{N} a_{t_i,n}p^d_{t_i,n} + l^d_{t_i}c^d_{t_i} + (E^+)^d_{t_i} - (E^-)^d_{t_i} = y^d_{t_i}, \forall d \in \mathbf{D}, \end{aligned} \tag{4.18}$$

$$(E^+)^d_{t_i}, (E^-)^d_{t_i} \geq 0, \forall d \in \mathbf{D}, \tag{4.19}$$

$$a_{t_i,n} \text{ unrestricted}, \forall n \in \mathbf{N}, \tag{4.20}$$

$$l_{min} \leq l^d_{t_i} \leq l_{max}, \forall d \in \mathbf{D}, \tag{4.21}$$

where $E^d_{t_i}$ denotes the error term at time slot $t_i$ on day $d$. As discussed in previous subsection, the error term at time slot $t_i$ on day $d$ ($E^d_{t_i}$) is split into $(E^+)^d_{t_i}$ and $(E^-)^d_{t_i}$ to estimate the positive and negative errors in measurement, respectively. Similarly, a lower $f$ is preferred for higher fraud detection accuracy. Based on the anomaly coefficient of consumer $n$ at time interval $t_i$ ($a_{t_i,n}$) and loss factor at time interval $t_i$ on day $d$ ($l^d_{t_i}$), the

University of Malaya

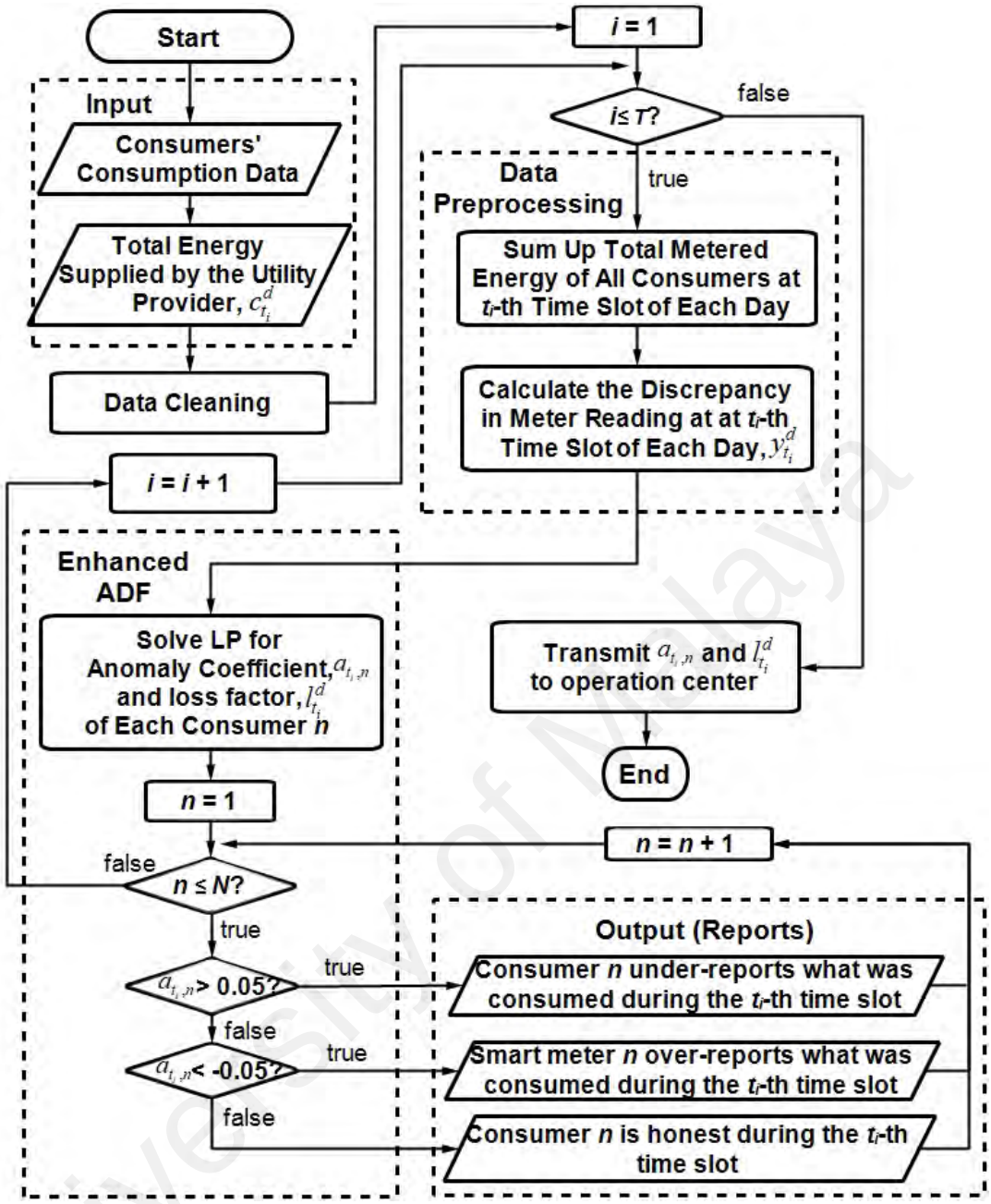



**Figure 4.3: Flow chart of the Enhanced ADF scheme.**

collector can pinpoint the location and period of energy diversions and/or faulty SMs as well as estimate the percentage of TLs. Flow chart in Figure 4.3 describes the operations in the Enhanced ADF scheme.

Similar to ADF, the collector labels the SM of all consumers in the NAN of interest from 1 to $N$. Before Enhanced ADF is invoked, data cleaning is performed to remove the rows of missing/corrupted data. The removal of those missing/corrupted data rows will not

University of Malaya

affect the anomaly detection analysis as long the number of observations are greater than the number of consumers in the service area. Correspondingly, the faulty or compromised SMs are inspected and replaced so that the Enhanced ADF can obtain a more precise NTL detection analysis. The scheme begins by forming a LSE according to time slot $t_i$, followed by computing the mismatches between the total supplied power and the sum of consumers' reported energy consumption of each day. Subsequently, $a_{t_i,n}$, $l_{t_i}^d$, $(E^+)_{t_i}^d$ and $(E^-)_{t_i}^d$ are solved using Equations (4.18)-(4.21).

For every consumer $n \in \mathbf{N}$, if the computed $a_{t_i,n} > 0.05$, it indicates that consumer $n$ tampers his SM to under-report the SM readings during the $t_i$-th time slot. Conversely, $a_{t_i,n} < -0.05$ indicates that the $n$-th SM is out of order and reports more than what was consumed during the $t_i$-th time slot. Otherwise, $a_{t_i,n} \in [-0.05, 0.05]$ implies that consumer $n$ is honest (i.e., $a_{t_i,n} \approx 0$), thereby the SM is neither fraudulent nor faulty during the $t_i$-th time interval. Finally, the collector will transmit the computed consumers' anomaly coefficient, $a_{t_i,n}$ and percentage of TLs of the $d$-th day, $l_{t_i}^d$ respectively, at the $t_i$-th time interval to the operation center. In the enhanced framework, the collector invokes Enhanced ADF after data collection has completed (i.e., at least $N$ days, where $N$ is the size of the service area) to observe the cheating patterns of each consumer. In this thesis, it is assumed that each dishonest consumer always attempts to steal energy when their load demand is higher. However, if the dishonest consumer has different cheating patterns during weekend/public holidays, the dataset during that period will be analyzed separately by doing minor modification to the proposed Enhanced ADF.

As discussed earlier in Section 4.4.1, the proposed anomaly detection framework detects NTLs based on the energy balance analysis. Specifically, the proposed framework shortlists areas with high probability of NTLs according to the discrepancy of meter readings at the DS and model the amount of stolen energy at a SM as an anomaly coefficient. However,

University of Malaya

the proposed framework cannot detect theft attack that evades the balance check. For instance, an energy thief who compromises a neighbor's SM or by physically tapping into the neighbor's electrical system to ensure that the consumption of at least one of his/her neighbors is over-reported will escape from the anomaly detection. In such a case, the innocent neighbor will pay for the energy thief's electricity.

### 4.5.4 Differences of Data Involved for ADF and Enhanced ADF

In this section, graphical illustrations to show the differences of data involved for the computation of ADF (i.e., Equation (4.6)) and Enhanced ADF (i.e., Equation (4.17)) are presented to elaborate each scheme for the beneficial of reader to understand the proposed LP-based schemes.

Consider a NAN consisting of $N$ consumers. Suppose that the consumers' SM readings are sampled over $T$ time intervals everyday for a period of $D$ days. Recall that $p^d_{t_i,n}$ denotes the energy consumption recorded by consumer $n$ on day $d \in \mathbf{D}$ at time interval $t_i \in \mathbf{T}$. Meanwhile, $c^d_{t_i}$ and $l^d_{t_i}$ denote the total energy supplied by the UPs and the loss factor, respectively, at time interval $t_i$ on day $d$. The discrepancy in meter reading at time slot $t_i$ on day $d$ is denoted by $y^d_{t_i}$.

As preliminary work, it is assumed that the fraudulent consumers never stop cheating and the defective SMs are out of order all the time. In other words, the rate and cheating pattern of under-reporting/over-reporting remain the same throughout the period of observation. That is, the anomaly coefficient for each consumer $a_n$ remain *constant* all the time. The computation to solve the LSE in Equation (4.6) for the detection of *constant* anomaly coefficients using the ADF scheme is illustrated in Figure 4.4. One-day half-hourly metered energy consumption data (i.e., highlighted in blue) are required for the anomaly detection analysis as long the number of observations is greater than the number of consumers (i.e., $T > N$) in the service area. For service area of larger size especially when number of



University of Malaya

**ADF**

| Day 1 | Consumer 1 | Consumer 2 | ... | Consumer *N* | Collector | Discrepancy in Meter Reading |
|---|---|---|---|---|---|---|
| 1st interval, $t_1$ | $a_1 p^1_{t_1,1}$ | $a_2 p^1_{t_1,2}$ | ... | $a_N p^1_{t_1,N}$ | $l^1_{t_1} c^1_{t_1}$ | $y^1_{t_1}$ |
| 2nd interval, $t_2$ | $a_1 p^1_{t_2,1}$ | $a_2 p^1_{t_2,2}$ | ... | $a_N p^1_{t_2,N}$ | $l^1_{t_2} c^1_{t_2}$ | $y^1_{t_2}$ |
| | | | ⋮ | | | |
| $T^{th}$ interval, $t_T$ | $a_1 p^1_{t_T,1}$ | $a_2 p^1_{t_T,2}$ | ... | $a_N p^1_{t_T,N}$ | $l^1_{t_T} c^1_{t_T}$ | $y^1_{t_T}$ |

Anomaly Detection for Day 1

| Day 2 | Consumer 1 | Consumer 2 | ... | Consumer *N* | Collector | Discrepancy in Meter Reading |
|---|---|---|---|---|---|---|
| 1st interval, $t_1$ | $a_1 p^2_{t_1,1}$ | $a_2 p^2_{t_1,2}$ | ... | $a_N p^2_{t_1,N}$ | $l^2_{t_1} c^2_{t_1}$ | $y^2_{t_1}$ |
| 2nd interval, $t_2$ | $a_1 p^2_{t_2,1}$ | $a_2 p^2_{t_2,2}$ | ... | $a_N p^2_{t_2,N}$ | $l^2_{t_2} c^2_{t_2}$ | $y^2_{t_2}$ |
| | | | ⋮ | | | |
| $T^{th}$ interval, $t_T$ | $a_1 p^2_{t_T,1}$ | $a_2 p^2_{t_T,2}$ | ... | $a_N p^2_{t_T,N}$ | $l^2_{t_T} c^2_{t_T}$ | $y^2_{t_T}$ |

Anomaly Detection for Day 2

| Day *D* | Consumer 1 | Consumer 2 | ... | Consumer *N* | Collector | Discrepancy in Meter Reading |
|---|---|---|---|---|---|---|
| 1st interval, $t_1$ | $a_1 p^D_{t_1,1}$ | $a_2 p^D_{t_1,2}$ | ... | $a_N p^D_{t_1,N}$ | $l^D_{t_1} c^D_{t_1}$ | $y^D_{t_1}$ |
| 2nd interval, $t_2$ | $a_1 p^D_{t_2,1}$ | $a_2 p^D_{t_2,2}$ | ... | $a_N p^D_{t_2,N}$ | $l^D_{t_2} c^D_{t_2}$ | $y^D_{t_2}$ |
| | | | ⋮ | | | |
| $T^{th}$ interval, $t_T$ | $a_1 p^D_{t_T,1}$ | $a_2 p^D_{t_T,2}$ | ... | $a_N p^D_{t_T,N}$ | $l^D_{t_T} c^D_{t_T}$ | $y^D_{t_T}$ |

Anomaly Detection for Day *D*

**Figure 4.4: Graphical illustration to show the data involved for the computation of the ADF scheme.**

consumers is greater than the number of observations (i.e., $T < N$), consumers' power consumption data are observed over longer period to increase the detection accuracy.

Nonetheless, it is observed that the ADF scheme may not be numerically stable when some of the fraudulent consumers steal energy inconsistently. Specifically, ADF may not detect all energy thieves when they cheat during random intervals in a day. In such a case, the anomaly coefficient for each consumer at time interval $t_i$, $a_{t_i,n}$ *varies* at each time interval. To overcome the deficiency of the ADF scheme, an Enhanced ADF scheme is put forward to reveal the locations and periods of intermittent energy theft or device failure. The computation to solve LSE in Equation (4.17) for the detection of *random* and

University of Malaya

*varying* anomaly coefficients using the Enhanced ADF scheme is graphically represented in Figure 4.5. In Enhanced ADF, the SM readings are analyzed over a longer period (i.e., at least $N$ days, where $N$ is the number of consumers in the service area) *according to specific time slot* $t_i \in \mathbf{T}$ of each day until the computed values converge. This is due to the fact that observation of metered data over longer periods leads to addition in the number of constraints that can improve the accuracy of the theft detection analysis. As shown in Figure 4.5, the LSE highlighted in orange formulates the anomaly detection for the first time interval $t_i$ of each day while the LSE highlighted in red shows the anomaly detection for the $T$-th interval $t_T$ of each day. In such a case, the $a_{t_i,n}$ and $l^d_{t_i}$ are solved according to specific time interval of each day to detect more sophisticated NTL attacks such as intermittent electricity pilfering and/or meter irregularities.

## 4.6 Summary of Chapter

This chapter puts forward two new anomaly detection schemes, namely ADF and Enhanced ADF, which takes into consideration the impact caused by TLs and measurement noise on NTL detection analysis, with the aim to improve the detection accuracy and minimize the number of false positives. The two proposed schemes are based on LP. The schemes shortlist service areas with high probability of theft according to the discrepancy of meter readings at the DS and model the amount of stolen energy at a SM as an anomaly coefficient. Similar to the previously proposed MLR-based energy theft detection schemes, any non-zero anomaly coefficients are indicative of energy frauds or metering irregularities. In addition, the proposed framework is also able to estimate the percentage of TLs based on measurements at the data collector and the knowledge of the distribution network. Nonetheless, it is observed that some of the energy thieves/defective meters stay undetected by the ADF scheme when electricity pilfering/metering defects occur only during a certain period in a day. To overcome the deficiency of the ADF scheme, the metered energy

University of Malaya


**Figure 4.5: Graphical illustration to show the data involved for the computation of the Enhanced ADF scheme.**

consumption data are observed over long periods and an Enhanced ADF scheme is put forward to solve more diverse and sophisticated attack types so that the proposed model can still detect meter irregularities even when there are intermittent faulty equipment/cheating.

In order to assess the performance of both the LR-based and LP-based anomaly detection frameworks, Matlab simulation and test rig-based experiments were performed. The data collection and test setup will be presented in the next chapter.

University of Malaya

# CHAPTER 5: DATA COLLECTION AND TEST SETUP

## 5.1 Overview

This chapter describes the data collection and test setup conducted prior the evaluation of the proposed anomaly detection frameworks. Two types of smart metering data from different sources are presented, whereby the details are elaborated in the subsections. The first subsection discusses the smart metering data extracted from the Smart Metering Electricity Trial, followed by the description of data extracted from the hardware experimentation in the laboratory. Then, these data are preprocessed and transformed into the required format for MLR and LP by performing *data cleaning and preprocessing* before the invocation of the proposed detection frameworks. The test setup of the proposed frameworks using Matlab is presented in the next section. The last section discusses the attack model.

## 5.2 Data Collection

Data collection is one of the significant stages in this thesis. Notably, the data collection for this research study is performed in two phases. In the first phase of data collection, the smart energy data are extracted from the Irish Smart Energy Trial in this study. The smart energy dataset consists of half-hourly energy consumption reports for both Irish *residential* and *commercial* premises of different contracted power during 2009 and 2010.

Initially, it is assumed that TLs in power line transmission are trivial and hence ignored. The data collector measurement is obtained by duly summing up the energy consumption of all consumers in the service area at each time interval when there are no energy thefts and defective SMs, viz., original untampered data. Since the real NTL data samples and the SM readings measured by data collector are non-existent because SG is not fully deployed in Malaysia, the NTL scenarios such as energy fraud and meter irregularities are

University of Malaya

realized by artificially tampering the SM readings in simulations, which will be discussed in Section 5.5.

In the second phase of data collection, an AMI test rig is designed and built in the laboratory to validate the reliability and performance of the proposed anomaly detection frameworks in real SG environment. The detailed description of the test rig will be discussed in Section 5.2.2. Similar data are recorded in comparison to the first phase of data collection. Apart from the half-hourly consumers' energy consumption data, the SM readings of total energy supplied by the UP to the service area (i.e., data collector measurement) are also collected from the test rig.

### 5.2.1 Smart Metering Data from the Irish Smart Energy Trial

As mentioned in previous section, the smart energy data from the Irish Smart Energy Trial are extracted in this research study. The dataset is released by Commission for Energy Regulation (2009) in January 2012. The dataset includes half-hourly energy consumption reports of over 5000 Irish homes and small businesses during 2009 and 2010. The dataset shows the maximum energy consumed during 30-minute interval (in kilowatt-hour (kWh)) are 10kWh and 30kWh for residential and commercial premises, respectively. Consumers who agreed to participate in the energy trial had a SM installed in their premise. Therefore, it is reasonable to assume that all data samples belong to honest consumers. The large number and variety of consumers, long period of measurements and availability to the public make this dataset an appropriate source for research in the area of data analytics.

Technically, the Irish Smart Energy Trial (Commission for Energy Regulation, 2009) consists of four main components:

1. **SM**: The SMs used in the energy trial are single-phase meters. These meters provide a range of functions including export, import and reactive power register readings,

University of Malaya

**Table 5.1: Description of consumers' energy consumption data extracted from the Irish Smart Energy Trial**

| Column | Column Title | Description |
|---|---|---|
| 1 | ID | Consumer's Smart Meter ID |
| 2 | Five Digit Code | Day Code (digits 1-3, whereby day 001 = 1st January 2009)<br>Time Code (digits 4-5, whereby 1-48 for each 30 minutes with 1= 00:00:00 – 00:29:59) |
| 3 | kWh | Energy consumed by each consumer during 30 minute interval (in kWh) |

half-hourly profiles and an embedded load-rated switch for remote operation. Besides that, event and alarm indications such as meter error, meter cover open, contract exceeded, over-voltage and etc. are also available on the meter.

2. **Neighborhood Area Network (NAN)**: Data collector manages the communications processes in NAN. The data collector is connected to the three phases and neutral of the LV side of the DT and communicates with each SM via Power Line Carrier (PLC) communications over the phase and neutral.
3. **Wide Area Network (WAN)**: The WAN communications between the data collectors and head end are managed through the Vodafone network. General Packet Radio Service (GPRS) modems are equipped in each data collector.
4. **Head End System**: The head end system is in charge of data collection. The head end performs automatic reading of the data collected by the data collectors and stores these data in the MDMS database.

The consumers' SM data extracted from the Irish Smart Energy Trial consists of six zipped files named 'File1.txt.zip' to 'File6.txt.zip', whereby each containing one data file. Each data file consists of half-hourly metering reports for a 535 day for each consumer. Each data file is arranged into three data columns as shown in Figure 5.1. Table 5.1 details the consumers' energy consumption data extracted from the Irish Smart Energy Trial.

Besides, another file named as "SME and Residential allocations" is also included to describe the category and tariff of each consumer. These allocation data are arranged into

University of Malaya

| ID | Five Digit Code | kWh |
|---|---|---|
| 1392 | 19503 | 0.14 |
| 1392 | 19504 | 0.138 |
| 1392 | 19505 | 0.14 |
| 1392 | 19506 | 0.145 |
| 1392 | 19507 | 0.145 |
| 1392 | 19501 | 0.157 |
| 1392 | 19502 | 0.144 |
| 1392 | 19724 | 0.128 |
| 1392 | 19725 | 0.142 |
| 1392 | 19726 | 0.145 |
| 1392 | 19727 | 0.149 |
| 1392 | 19728 | 0.131 |
| 1392 | 19729 | 0.13 |
| 1392 | 19730 | 0.13 |
| 1392 | 19731 | 0.152 |
| 1392 | 19732 | 0.143 |
| 1392 | 19733 | 0.13 |
| 1392 | 19734 | 0.132 |
| 1392 | 19735 | 0.169 |
| 1392 | 19736 | 0.254 |
| 1392 | 19737 | 0.231 |
| 1392 | 19738 | 0.371 |
| 1392 | 19739 | 0.266 |

**Figure 5.1: The screen shot of consumers' energy consumption data extracted from the Irish Smart Energy Trial.**

**Table 5.2: Consumers' allocation information extracted from the Irish Smart Energy Trial**

| Column | Column Title | Description |
|---|---|---|
| 1 | ID | Consumer's Smart Meter ID |
| 2 | Code | The category of each consumer (i.e., 1= Residential, 2=SME and 3=others) |
| 3 | Residential Stimulus | The stimulus code for residential consumers (refer to Figure 5.2 for details) |
| 4 | Residential Tariff | The tariff category for residential consumers (refer to Figure 5.2 for details) |
| 5 | SME Allocation | The stimulus for SME consumers (refer to Figure 5.2 for details) |

five data columns as shown in Figure 5.2. Table 5.2 describes the consumers' allocation information extracted from the Irish Smart Energy Trial.

University of Malaya

| ID | Code | Residential - Tariff allocation | Residential - stimulus allocation | SME allocation |
|---|---|---|---|---|
| 1000 | 3 | | | |
| 1001 | 3 | | | |
| 1002 | 1 | E | E | |
| 1003 | 1 | A | 4 | |
| 1004 | 1 | A | 2 | |
| 1005 | 1 | D | 4 | |
| 1006 | 1 | C | 3 | |
| 1007 | 3 | | | |
| 1008 | 1 | D | 1 | |
| 1009 | 1 | A | 2 | |
| 1010 | 3 | | | |
| 1011 | 3 | | | |
| 1012 | 3 | | | |
| 1013 | 1 | D | 4 | |
| 1014 | 1 | E | E | |
| 1015 | 1 | C | 3 | |
| 1016 | 1 | B | 2 | |
| 1017 | 3 | | | |
| 1018 | 1 | E | E | |
| 1019 | 3 | | | |
| 1020 | 1 | B | 4 | |
| 1021 | 2 | | | C |
| 1022 | 1 | E | E | |
| 1023 | 2 | | | 2 |
| 1024 | 1 | C | 2 | |
| 1025 | 1 | C | 1 | |
| 1026 | 2 | | | 3 |

**Code**

| | |
|---|---|
| 1 | Residential |
| 2 | SME |
| 3 | Other |

**Residential Stimulus**

| | |
|---|---|
| E | Control |
| 1 | Bi-monthly detailed bill |
| 2 | Monthly detailed bill |
| 3 | Bi-monthly detailed bill +IHD |
| 4 | Bi-monthly detailed bill +OLR |
| W | Weekend tariff |

**Residential Tariff**

| | |
|---|---|
| E | Control |
| A | Tariff A |
| B | Tariff B |
| C | Tariff C |
| D | Tariff D |
| W | Weekend tariff |

**SME**

| | |
|---|---|
| 1 | Monthly detailed bill |
| 2 | Bi-monthly detailed bill +IOD |
| 3 | Bi-monthly detailed bill +web-access |
| 4 | Bi-monthly detailed bill |
| C | Control |

**Figure 5.2: The screen shot of consumers' allocation information extracted from the Irish Smart Energy Trial.**

### 5.2.2 Smart Metering Data from the Hardware Experimentation

In the second phase of data collection, an AMI test rig, consisting of three consumers, an operation center and a DS as illustrated in Figure 5.3 is designed and constructed in the laboratory as shown in Figure 5.4 to evaluate the performance of the proposed anomaly detection frameworks in real SG environment. The schematic diagram of the AMI test rig is designed as shown in Figure 5.5 so that each consumer can select different loads at each time interval to simulate real-world load profiles. For safety purpose, a rectangular-shape metal box known as the "load bank" shown in Figure 5.6 is designed to place all the resistive loads. Each layer of resistive loads represents the load of each consumer. Referring to the schematic diagram in Figure 5.5, three single-phase SMs are used to record the energy consumption of each consumer. A master SM, known as the *data collector*, is endowed in the DS to track the total power supplied by the UP (i.e., three-phase power supply) to all three consumers at each time interval. Then, all the SMs as well as the data collector are configured to send their energy consumption readings to the operation center (i.e., Omron NJ101-1020 controller (Omron, 2017)) at half-hourly interval. The

University of Malaya

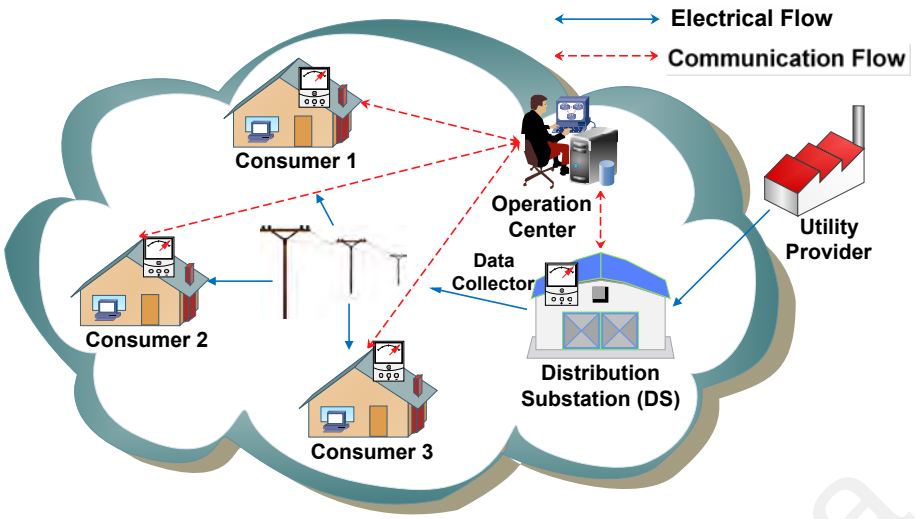



**Figure 5.3: The design of an AMI test rig in the laboratory.**

`rand` function packaged in Matlab R2014b is utilized to generate random load demand for each consumer at every time interval. Subsequently, the randomly generated load demand for each consumer is varied through the Miniature Circuit Breakers (MCBs) in the test rig to simulate real-world load profiles. Polyvinyl chloride (PVC) insulated 10mm$^2$ LV distribution copper cables are used to connect the DS and consumers. The cable length between two consumers and the cable length from the DS to the first consumer are 7m and 8.23m, respectively.

#### 5.2.2.1 Phoenix Series 2 Single-phase Smart Meter

SM is one of the key components in AMI. Real-time recording and monitoring of energy consumption can be made possible through the deployment of SMs. These digitized devices are progressively deployed to replace its antiquated predecessors to measure and monitor consumers' energy consumption in SGs. Transmission Control Protocol/Internet Protocol (TCP/IP) is utilized by these modernized meters to communicate with the UPs. Besides, they also help in mitigating NTLs as any attempts at meter tampering will be detected on the spot. Aside from collating information on power quality and recording energy consumption, the firmware of SMs can also be upgraded automatically over the

University of Malaya

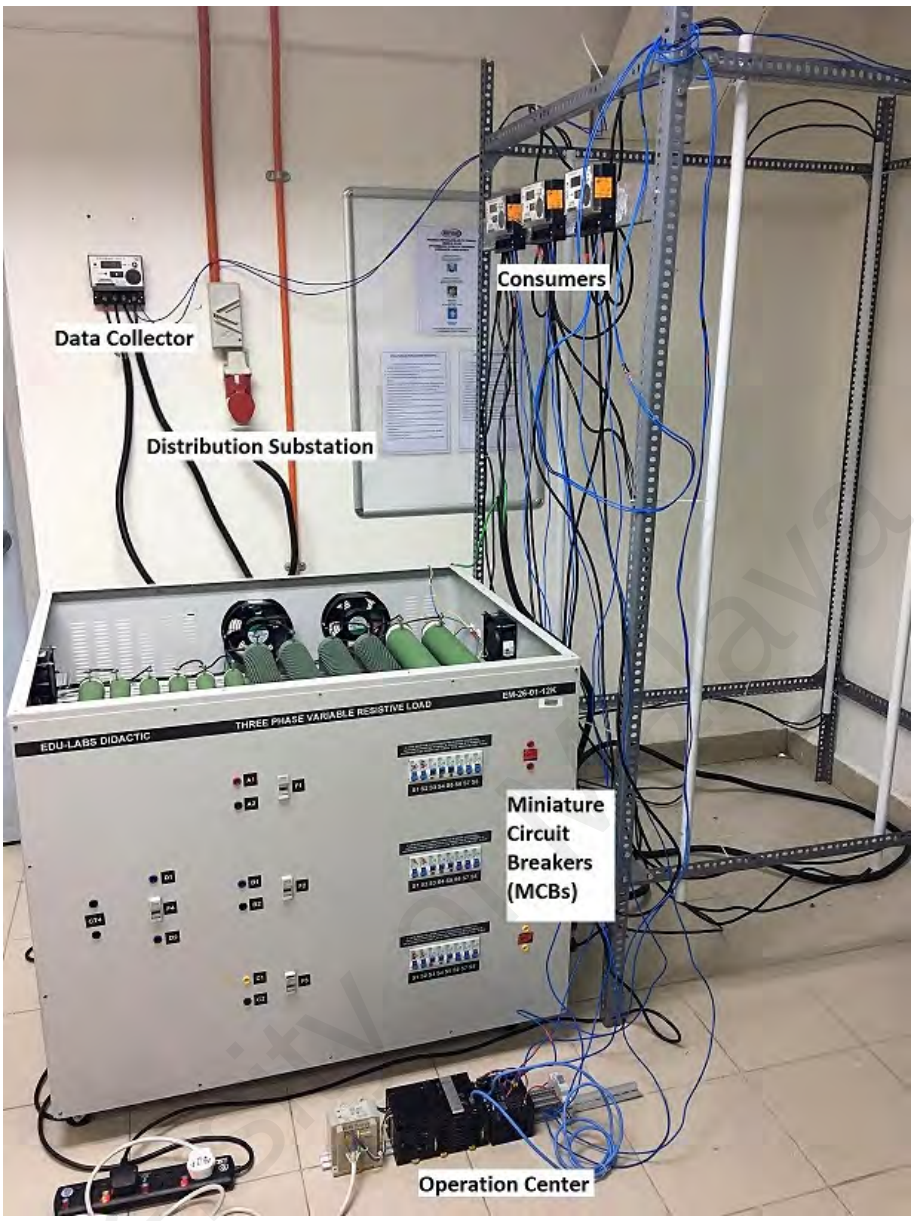



**Figure 5.4: The hardware experimentation of the AMI test rig.**

internet. Also, they can detect surges and outage events in consumers' premises. Therefore, the Phoenix Series 2 SM as shown in Figure 5.7, is rolled-out by TNB, Malaysia's largest UP, in the pilot SM project in Malacca and Putrajaya (Energy Commission, 2015). The same SM model is used in the AMI test rig to track the energy consumption of each consumer as well as to measure the aggregated power supply from the UP. It is a class 2 energy meter, mainly applicable for domestic with the current rating of 10A-100A.

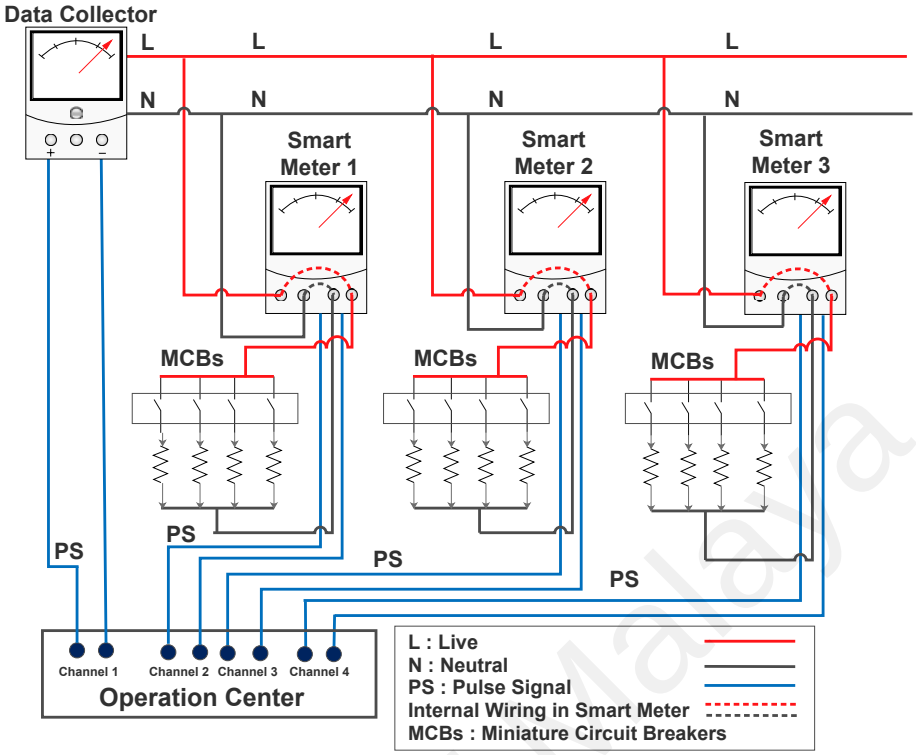



**Figure 5.5: The schematic diagram of the AMI test rig.**

**Figure 5.6: The load bank which contains all the resistive loads.**

University of Malaya


**Figure 5.7: The Phoenix Series 2 smart meter.**

**Figure 5.8: Omron NJ101-1020 machine automation controller powered on by Omron S82K-05024 power supply, serves as the operation center in the test rig.**

#### 5.2.2.2 Data Logger

In this thesis, Omron NJ101-1020 machine automation controller (Omron, 2017) powered on by Omron S82K-05024 power supply as demonstrated in Figure 5.8, is used as data logger to track the energy consumption data of each consumer as well as to record the aggregated power supplied by the UP at the predefined time interval (i.e., 30 minutes interval). The data logger acts as the operation center in the test rig.

Universiti Malaya

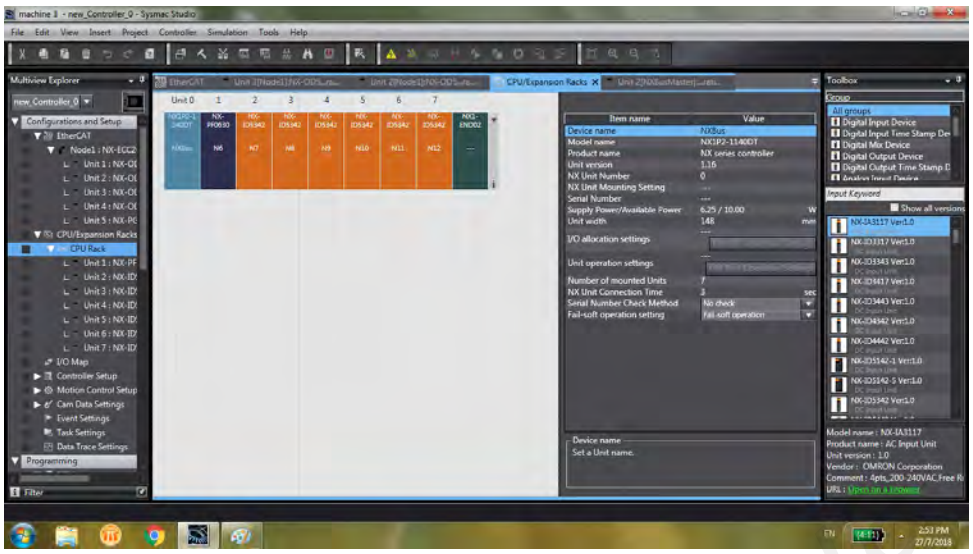


**Figure 5.9: Configuration page to setup the Omron controller.**

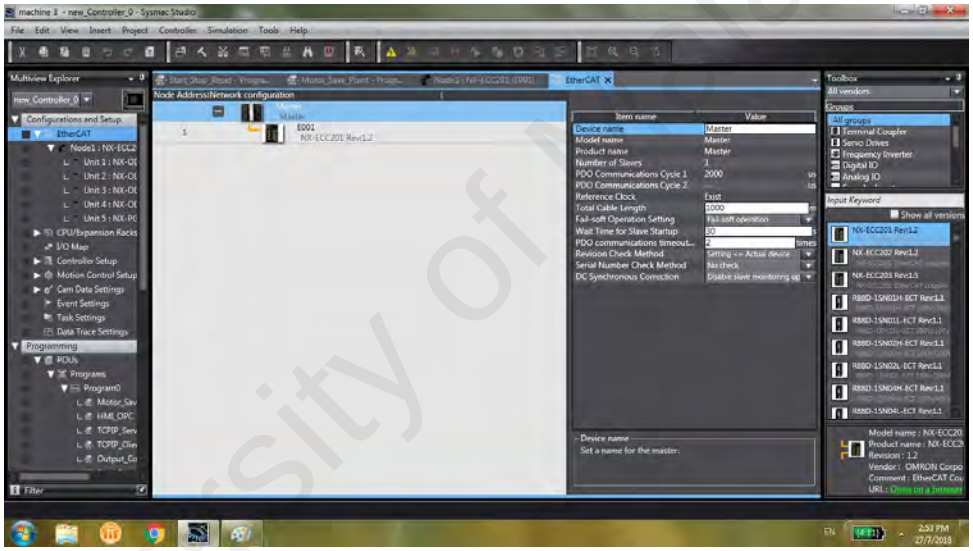


**Figure 5.10: Network configuration page to setup the communications between controller modules.**

To program the data logging function of Omron controller, Sysmac Studio Automation Software Version 1.21 is used. This software offers an integrated development environment for Omron NJ-series controllers. Users can perform debugging and testing of logic, safety, motion and vision sensors, integrate programming and achieve an advanced security function with the user interface. Figures 5.9, 5.10 and 5.11 show the configuration and programming pages of the software to setup the data logging function of the Omron controller.

University of Malaya

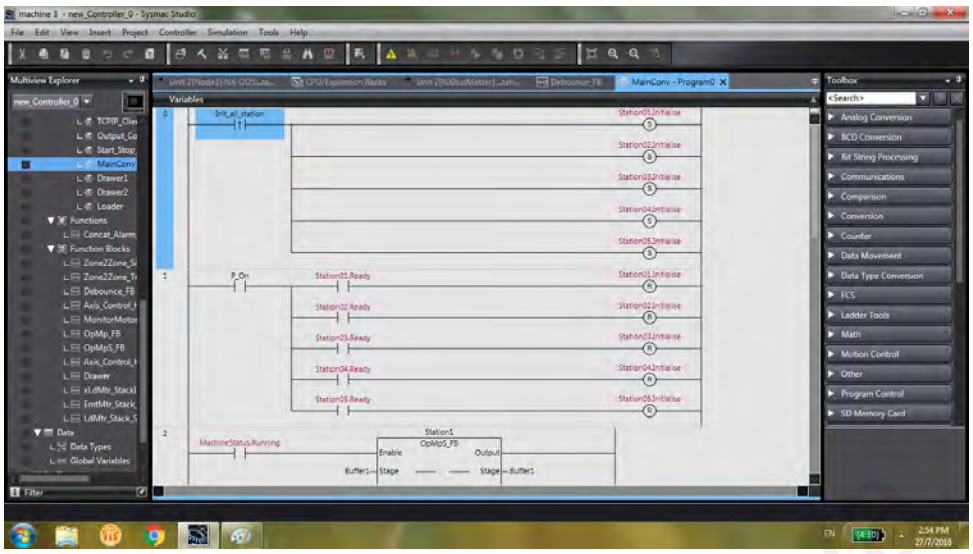


**Figure 5.11: Programming page to setup the function of each module.**


**Figure 5.12: Miniature circuit breakers are attached to a metal plate at the front part of the load bank to prevent electrical shock.**

### 5.2.2.3 Miniature Circuit Breaker

Miniature Circuit Breakers (MCBs) are usually used for short circuit or overload protection in the distribution system. As mentioned in Section 5.2.2, the randomly generated load demand for each consumer is varied through the MCBs in the test rig to simulate real-world load profiles. For safety purpose, all the MCBs are attached to a metal plate at the front part of the load bank as shown in Figure 5.12 to prevent electrical shock.

University of Malaya

## 5.3 Data Cleaning and Preprocessing

In data analytic, missing data or unusual patterns caused by unplanned events or the failure of data is known as *bad data* (Y. Wang, Chen, Hong, & Kang, 2018). Meanwhile, data cleaning is the process of identifying inaccurate, corrupted records and bad data from a dataset, table, or database and then replacing or deleting the dirty or coarse data (Wu, 2013). In this work, data cleaning is performed to filter out suspicious SMs which report a constant low value, do not have measurements (i.e., 0 kWh) throughout the periods of observation or have corrupted data after data collection Ahmad et al. (2018); Dos Angelos et al. (2011). Correspondingly, the malicious or defective SMs are inspected and replaced so that the proposed anomaly detection frameworks can obtain a more accurate NTL detection analysis. Subsequently, the energy consumption data are extracted and transformed into the required format for MLR and LP, as discussed in Sections 3.6.3 and 4.5.4, by performing data preprocessing. The collected data are represented by their 30-minute consumption profiles. Figure 5.13 illustrates the half-hourly kWh energy consumption sample data which are extracted and transformed into the required format for *regression analysis*, whereby the energy consumption data for 15 consumers are represented by data columns B through P. The aggregated power supplied by the UPs at each interval (i.e., data collector measurement) is represented by data column Q. Data column A depicts the "Five Digit Code" which represents the day and time interval. Recall that the day code is depicted by the first 3 digits (e.g., day 001 = 1st January 2009) and the time code is represented by the forth and fifth digits (e.g., 1-48 for each 30 minutes with 1= 00:00:00 – 00:29:59). Meanwhile, Figure 5.14 presents the 30-minute kWh energy consumption sample data that are extracted and transformed into the required format for *optimization analysis*. These dataset has additional two data columns in comparison to the transformed data for regression analysis, as illustrated in Figure 5.13. Specifically, two additional data columns

University of Malaya

(i.e., data columns R and S) are added for capturing the positive and negative calibration errors/noise of the equipment.

Both the detection schemes commence by computing the mismatches between the total metered energy consumption of all consumers in the service area and the total power supplied by the UPs at each time interval. Then, the detection and estimation are undertaken by the proposed LR-based and LP-based anomaly detection frameworks. Based on the regression and optimization results from Matlab, the UPs can easily identify the positions of energy fraud and meter irregularities. The test setup of the anomaly detection frameworks using Matlab will be discussed in the next section.

## 5.4 Test Setup

In this thesis, Matlab R2014b (MathWorks, 2017) is used for developing the anomaly detection frameworks. Matlab combines design processes with a programming language which expresses arrays and matrices directly and a desktop environment tuned for iterative analysis. After preprocessing the consumers' energy consumption data extracted from the Irish Smart Energy Trial into the required format, the `fitlm` function packaged in the Statistics Toolbox of Matlab R2014b is used to solve for the estimated anomaly coefficients $a_n$ in Equation (3.5) using MLR. The indicator for the constant intercept in the fit (i.e., $\alpha$ in Equation (3.8)) is configured as 'false' so that the response is fully dependent on the predictors $\mathbf{P}$. Next, the $a_n$, $t$-statistics and $p$-values of all consumers (i.e., $\forall n \in \mathbf{N}$) are retrieved from the LR analysis. Based on the estimated $a_n$ and its corresponding $p$-value, the locations of *consistent* energy frauds and/or faulty SMs can be pinpointed accurately. Nonetheless, the proposed LR-ETDM scheme might be unstable when there are *inconsistent* energy thefts and/or defective SMs. To overcome the deficiency of LR-ETDM, CVLR-ETDM is put forward by incorporating detection coefficients and categorical variables into MLR so that the scheme can successfully detect consumers' malfeasance

University of Malaya

and faulty meters even when there are varying cheating trends/meter irregularities, either during off-peak or on-peak period.

For the LP-based anomaly detection framework, the `linprog` function built in the Optimization Toolbox of Matlab R2014b is used to solve for $a_n$ and $l_{t_i}$ in Equation (4.11), and both $a_{t_i,n}$ and $l_{t_i}^d$ in Equation (4.18) by using either the dual-simplex method (Koberstein, 2008) or the interior point method (Boyd & Vandenberghe, 2004) of LP based on the fine-grained consumers' metered energy consumption readings. Similar to the LR-based energy theft detection framework, any non-zero anomaly coefficients are indicative of energy frauds or metering irregularities. Aside from detecting NTL events, the proposed framework is also able to estimate the percentage of TLs through loss factor based on measurements at the data collector and the knowledge of the distribution network. Besides that, the proposed detection framework is also capable of detecting under-reporting/over-reporting by SMs even when there are intermittent cheating and/or faulty equipment, and not restricted to detection during off-peak and on-peak periods only.

Furthermore, an AMI test rig is built to validate the performance and reliability of the proposed anomaly detection frameworks in real SG environment. A series of tests are conducted on the energy consumption data collected from the test rig. The test results obtained are correlated with Matlab simulation results in order to confirm the performance and reliability of the proposed anomaly detection frameworks in real world.

## 5.5 Attack Model

In this thesis, it is assumed that adversaries are the fraudulent consumers who compromise their SMs to fabricate energy readings to reduce their billings. Their goal is to under-report energy consumption and make monetary profit at the expense of the UPs.

As discussed in Section 2.3.2, there are various known techniques for stealing energy from the power grids. Recall that these energy fraud techniques, including those that

University of Malaya

are widely implemented in both conventional power grids and SGs, may be grouped into three categories, namely physical attacks, cyber attacks and data attacks as discussed in (McLaughlin et al., 2013). Note that data attacks could also be realized through threats from the cyber and physical attacks.

In the attack model, it is assumed that the SMs are in one of the three states, namely, honest, compromised or faulty. Suppose the SMs are configured to record benign half-hourly meter readings $\mathbf{P}_{\mathbf{n}}^{*} = \{p_{t_1,n}^{*}, p_{t_2,n}^{*}, \cdots, p_{t_{48},n}^{*}\}$, for time interval $t_i = t_1, t_2, \cdots, t_{48}$ (i.e., 48 data points in a day). In this thesis, the energy frauds or meter irregularities are simulated by tampering the benign SM readings $\mathbf{P}_{\mathbf{n}}^{*}$, as the real NTL data samples are non-existent in Malaysia because SG is not fully implemented. Let $p_{t_i,n}$ denote the energy consumption recorded by the $n$-th SM after the application of one of the state functions in Table 5.3. The possible states of the $n$-th SM, where $n \in \mathbf{N} = \{1, 2, \cdots, N\}$ and the corresponding types of energy fraud, are summarized in Table 5.3.

Particularly, the consumer is honest in $s_1$ as he/she reports the actual meter readings. In $s_2$, the SM readings are scaled by a constant percentage (i.e., constant rate). In other words, the fraudulent consumer either reports a fraction of his/her consumed energy consistently (e.g., $\nu \in (0, 0.95)$) or the SM over-reports the consumption readings all the time (e.g., $\nu \in (1.05, 2.5]$). In $s_3$, the energy thief pilfers energy only during certain periods in a day. For instance, the fraudulent commercial consumer reports 40% of the actual consumed data during operation hours (i.e., $\delta_{t_i} = 0.4$) and reports the actual consumption data at night (i.e., $\delta_{t_i} = 1$). Using the state function in $s_4$, the SM sends zero reading or does not have measurements (i.e., $\eta_{t_i} = 0$) only during certain periods in a day. Last but not least, $s_5$ reports the average of SM readings over the day. In this thesis, without loss of generality, faulty SMs are assumed to always report more than what the corresponding consumers actually consumed (i.e., $\nu \in (1.05, 2.5]$).

University of Malaya

| Five Digit Code | SM1 | SM2 | SM3 | SM4 | SM5 | SM6 | SM7 | SM8 | SM9 | SM10 | SM11 | SM12 | SM13 | SM14 | SM15 | Collector |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19501 | 0.175 | 0.37 | 0.39 | 0.403 | 0.34 | 0.33 | 0.17 | 0.37 | 0.37 | 0.42 | 0.66 | 0.48 | 0.432 | 1.12 | 0.59 | 7.89 |
| 19502 | 0.18 | 0.4 | 0.38 | 0.403 | 0.33 | 0.33 | 0.175 | 0.38 | 0.38 | 0.47 | 0.735 | 0.51 | 0.848 | 2.17 | 1.115 | 11.21 |
| 19503 | 0.25 | 0.54 | 0.49 | 0.442 | 0.35 | 0.36 | 0.215 | 0.46 | 0.45 | 0.72 | 1.065 | 0.72 | 0.984 | 2.49 | 1.275 | 13.57 |
| 19504 | 0.49 | 1.05 | 0.93 | 0.559 | 0.46 | 0.45 | 0.335 | 0.74 | 0.74 | 1.42 | 2.1 | 1.41 | 0.996 | 2.52 | 1.29 | 18.27 |
| 19505 | 0.545 | 1.16 | 1.01 | 0.624 | 0.51 | 0.49 | 0.375 | 0.83 | 0.81 | 1.65 | 2.415 | 1.62 | 0.968 | 2.45 | 1.245 | 19.37 |
| 19506 | 0.54 | 1.14 | 0.98 | 0.624 | 0.47 | 0.46 | 0.36 | 0.79 | 0.77 | 1.66 | 2.49 | 1.65 | 0.936 | 2.37 | 1.195 | 18.96 |
| 19507 | 0.495 | 1.03 | 0.9 | 0.585 | 0.46 | 0.45 | 0.335 | 0.72 | 0.71 | 1.52 | 2.31 | 1.51 | 0.864 | 2.2 | 1.105 | 17.52 |
| 19508 | 0.465 | 0.98 | 0.82 | 0.598 | 0.48 | 0.47 | 0.33 | 0.71 | 0.7 | 1.4 | 2.1 | 1.39 | 0.768 | 1.94 | 0.98 | 16.22 |
| 19509 | 0.425 | 0.88 | 0.71 | 0.559 | 0.45 | 0.43 | 0.315 | 0.68 | 0.64 | 1.3 | 1.935 | 1.27 | 0.708 | 1.76 | 0.895 | 14.88 |
| 19510 | 0.375 | 0.76 | 0.63 | 0.546 | 0.44 | 0.39 | 0.285 | 0.61 | 0.58 | 1.16 | 1.695 | 1.14 | 0.664 | 1.66 | 0.84 | 13.58 |
| 19511 | 0.35 | 0.69 | 0.57 | 0.559 | 0.43 | 0.4 | 0.275 | 0.59 | 0.54 | 1.07 | 1.575 | 1.03 | 0.616 | 1.52 | 0.775 | 12.66 |
| 19512 | 0.38 | 0.74 | 0.63 | 0.65 | 0.52 | 0.46 | 0.32 | 0.65 | 0.6 | 1.06 | 1.545 | 0.98 | 0.588 | 1.47 | 0.72 | 12.95 |
| 19513 | 0.435 | 0.85 | 0.7 | 0.741 | 0.55 | 0.5 | 0.345 | 0.66 | 0.6 | 1.03 | 1.44 | 0.92 | 0.588 | 1.42 | 0.69 | 13.17 |
| 19514 | 0.44 | 0.8 | 0.63 | 0.754 | 0.52 | 0.48 | 0.36 | 0.66 | 0.6 | 1.07 | 1.395 | 0.88 | 0.576 | 1.35 | 0.64 | 12.82 |
| 19515 | 0.47 | 0.82 | 0.62 | 0.806 | 0.52 | 0.49 | 0.385 | 0.69 | 0.58 | 1.14 | 1.41 | 0.89 | 0.54 | 1.18 | 0.545 | 12.64 |
| 19516 | 0.465 | 0.84 | 0.66 | 0.832 | 0.53 | 0.49 | 0.375 | 0.65 | 0.58 | 1.06 | 1.365 | 0.85 | 0.392 | 0.82 | 0.37 | 11.43 |
| 19517 | 0.405 | 0.78 | 0.7 | 0.741 | 0.55 | 0.53 | 0.31 | 0.62 | 0.6 | 0.87 | 1.35 | 0.82 | 0.3 | 0.71 | 0.33 | 10.49 |
| 19518 | 0.28 | 0.62 | 0.59 | 0.611 | 0.5 | 0.51 | 0.235 | 0.54 | 0.54 | 0.59 | 1.005 | 0.68 | 0.244 | 0.67 | 0.33 | 8.68 |
| 19519 | 0.23 | 0.53 | 0.57 | 0.559 | 0.47 | 0.53 | 0.205 | 0.49 | 0.54 | 0.46 | 0.84 | 0.6 | 0.22 | 0.65 | 0.355 | 7.96 |
| 19520 | 0.21 | 0.53 | 0.6 | 0.533 | 0.48 | 0.55 | 0.195 | 0.46 | 0.54 | 0.42 | 0.735 | 0.58 | 0.204 | 0.65 | 0.365 | 7.76 |
| 19521 | 0.2 | 0.49 | 0.61 | 0.507 | 0.44 | 0.56 | 0.185 | 0.46 | 0.57 | 0.41 | 0.735 | 0.61 | 0.192 | 0.61 | 0.385 | 7.66 |
| 19522 | 0.195 | 0.48 | 0.61 | 0.507 | 0.45 | 0.57 | 0.185 | 0.45 | 0.57 | 0.4 | 0.72 | 0.62 | 0.188 | 0.61 | 0.38 | 7.62 |
| 19523 | 0.195 | 0.5 | 0.62 | 0.481 | 0.43 | 0.58 | 0.18 | 0.43 | 0.59 | 0.38 | 0.765 | 0.66 | 0.184 | 0.61 | 0.405 | 7.7 |
| 19524 | 0.2 | 0.5 | 0.69 | 0.533 | 0.48 | 0.64 | 0.19 | 0.5 | 0.64 | 0.42 | 0.825 | 0.7 | 0.192 | 0.65 | 0.425 | 8.29 |

**Figure 5.13: The half-hourly kWh energy consumption sample data for the size of 15 consumers are extracted and transformed into the required format for regression analysis.**

Universiti Utara Malaysia

| Five Digit Code | SM1 | SM2 | SM3 | SM4 | SM5 | SM6 | SM7 | SM8 | SM9 | SM10 | SM11 | SM12 | SM13 | SM14 | SM15 | Collector | E+ | E- |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19501 | 0.1866 | 0.4 | 0.009 | 6.6079 | 0.822 | 0.063 | 1.3175 | 0.172 | 0.173 | 0.115 | 0.509 | 0.239 | 0.6855 | 0.647 | 0.0276 | 12.2128 | 1 | -1 |
| 19502 | 0.1578 | 0.413 | 0.056 | 6.4038 | 0.638 | 0.129 | 1.32 | 0.138 | 0.107 | 0.162 | 0.394 | 0.265 | 0.567 | 1.063 | 0.0544 | 12.32111 | 1 | -1 |
| 19503 | 0.1824 | 0.339 | 0.053 | 7.046 | 0.629 | 0.103 | 0.8555 | 0.145 | 0.082 | 0.096 | 0.478 | 0.241 | 0.375 | 0.602 | 0.0432 | 11.11337 | 1 | -1 |
| 19504 | 0.1836 | 0.309 | 0.008 | 6.825 | 0.543 | 0.061 | 0.893 | 0.157 | 0.191 | 0.162 | 0.471 | 0.267 | 0.1875 | 0.115 | 0.0452 | 10.15586 | 1 | -1 |
| 19505 | 0.1626 | 0.296 | 0.02 | 6.0424 | 0.163 | 0.127 | 0.866 | 0.126 | 0.086 | 0.095 | 0.326 | 0.252 | 0.1875 | 0.109 | 0.1132 | 8.928967 | 1 | -1 |
| 19506 | 0.1662 | 0.266 | 0.087 | 5.707 | 0.137 | 0.095 | 0.865 | 0.162 | 0.083 | 0.164 | 0.189 | 0.266 | 0.1875 | 0.109 | 0.0432 | 8.534626 | 1 | -1 |
| 19507 | 0.1422 | 0.285 | 0.008 | 6.0554 | 0.18 | 0.087 | 0.979 | 0.531 | 0.193 | 0.095 | 0.193 | 0.272 | 0.189 | 0.108 | 0.0392 | 9.517911 | 1 | -1 |
| 19508 | 0.1812 | 0.263 | 0.008 | 5.1844 | 0.187 | 0.105 | 0.897 | 0.14 | 0.082 | 0.164 | 0.182 | 0.252 | 0.1875 | 0.108 | 0.0584 | 8.096726 | 1 | -1 |
| 19509 | 0.1674 | 0.272 | 0.062 | 5.1714 | 0.115 | 0.083 | 0.853 | 2.701 | 0.082 | 0.095 | 0.091 | 0.276 | 0.144 | 0.108 | 0.0276 | 10.43678 | 1 | -1 |
| 19510 | 0.1362 | 0.306 | 0.047 | 5.8747 | 0.115 | 0.106 | 0.8525 | 2.956 | 2.122 | 0.163 | 0.134 | 0.25 | 0.0255 | 0.108 | 0.0784 | 13.51423 | 1 | -1 |
| 19511 | 0.1908 | 0.235 | 0.008 | 6.045 | 0.142 | 0.098 | 1.006 | 2.971 | 0.235 | 0.095 | 0.207 | 0.267 | 0.0675 | 0.107 | 0.0888 | 12.23543 | 1 | -1 |
| 19512 | 0.1524 | 0.235 | 0.011 | 5.6238 | 0.193 | 0.085 | 0.8625 | 2.989 | 0.191 | 0.164 | 0.196 | 0.132 | 0.2025 | 0.106 | 0.0368 | 11.17335 | 1 | -1 |
| 19513 | 0.1566 | 0.6 | 0.087 | 6.5026 | 0.129 | 0.12 | 0.8585 | 2.958 | 0.3 | 0.128 | 0.168 | 0.15 | 0.1905 | 0.111 | 0.0656 | 12.38225 | 1 | -1 |
| 19514 | 0.1638 | 0.24 | 0.425 | 6.0333 | 0.109 | 0.102 | 0.8615 | 2.948 | 0.318 | 0.182 | 0.197 | 0.149 | 0.1095 | 0.105 | 0.0316 | 12.03507 | 1 | -1 |
| 19515 | 0.1662 | 0.313 | 0.165 | 5.7759 | 0.115 | 0.383 | 1.5005 | 2.901 | 0.145 | 0.092 | 0.19 | 1.815 | 0.039 | 0.105 | 0.0344 | 14.78807 | 1 | -1 |
| 19516 | 0.1932 | 0.356 | 0.451 | 8.2134 | 0.192 | 0.13 | 1.319 | 3.305 | 0.157 | 0.162 | 0.44 | 4.961 | 1.4205 | 0.104 | 0.0904 | 21.57741 | 1 | -1 |
| 19517 | 0.1644 | 0.584 | 0.131 | 12.5255 | 0.187 | 4.05 | 1.54 | 2.903 | 0.498 | 0.092 | 0.41 | 3.956 | 2.445 | 0.103 | 0.0712 | 28.88142 | 1 | -1 |
| 19518 | 0.1752 | 0.424 | 0.116 | 11.9873 | 0.115 | 0.301 | 2.846 | 2.861 | 1.753 | 0.158 | 0.382 | 3.759 | 0.06 | 0.103 | 0.0512 | 26.41324 | 1 | -1 |
| 19519 | 0.1164 | 0.228 | 0.494 | 11.5102 | 0.115 | 0.171 | 2.9325 | 2.908 | 0.194 | 0.105 | 1.877 | 3.856 | 0.201 | 0.102 | 0.0492 | 26.29207 | 1 | -1 |
| 19520 | 0.1728 | 0.257 | 0.305 | 11.8092 | 0.364 | 0.148 | 2.9425 | 2.909 | 0.086 | 0.255 | 3.02 | 3.753 | 0.417 | 0.176 | 0.0272 | 28.29314 | 1 | -1 |
| 19521 | 0.1656 | 0.484 | 0.174 | 12.2161 | 0.303 | 0.1 | 2.906 | 2.407 | 0.029 | 0.508 | 0.958 | 3.718 | 0.7125 | 0.382 | 0.0604 | 25.95878 | 1 | -1 |
| 19522 | 0.1278 | 2.054 | 0.169 | 13.5122 | 0.158 | 0.127 | 2.7835 | 0.144 | 0.077 | 0.483 | 0.559 | 4.289 | 0.3555 | 0.245 | 0.0736 | 25.94916 | 1 | -1 |
| 19523 | 0.1596 | 1.282 | 0.155 | 14.0491 | 0.526 | 0.132 | 2.866 | 0.159 | 0.101 | 0.821 | 4.047 | 3.759 | 0.2385 | 0.827 | 0.0684 | 29.83833 | 1 | -1 |
| 19524 | 0.1512 | 0.264 | 0.114 | 13.7722 | 0.332 | 0.09 | 2.982 | 0.126 | 0.03 | 0.534 | 0.446 | 3.763 | 0.3135 | 1.243 | 0.068 | 25.39444 | 1 | -1 |

**Figure 5.14: The half-hourly kWh energy consumption sample data for the size of 15 consumers are extracted and transformed into the required format for optimization analysis.**

**Table 5.3: Possible states of the smart meters**

| State | State Function | Description | Type of Energy Fraud/Defect |
|---|---|---|---|
| $s_1$ | $p_{t_i,n} = 1 \times p^*_{t_i,n}$ | SM is neither compromised nor faulty | (a) Un-tampered connection |
| $s_2$ | $p_{t_i,n} = \nu p^*_{t_i,n}; \nu \in (0, 0.95) \cup (1.05, 2.5]$ | SM is compromised/faulty all the time | (a) Place strong magnet near meter<br>(b) Wire partial bypass of the meter<br>(c) Hack meter to falsify reading |
| $s_3$ | $p_{t_i,n} = \delta_{t_i} p^*_{t_i,n};$<br>$\delta_{t_i} = \begin{cases} \nu & \text{if } start \le t_i \le end \\ 1 & \text{otherwise} \end{cases}$<br>where $\nu$ is as defined in $s_2$ above; | SM is compromised/faulty<br>only during a certain period in a day | (a) Place strong magnet near meter<br>(b) Wire partial bypass of the meter<br>(c) Hack meter to falsify reading |
| $s_4$ | $p_{t_i,n} = \eta_{t_i} p^*_{t_i,n};$<br>$\eta_{t_i} = \begin{cases} 0 & \text{if } start \le t_i \le end \\ 1 & \text{otherwise} \end{cases}$ | SM sends zero reading<br>only during a certain period in a day | (a) Complete meter bypass<br>(b) Direct connection to the grid<br>(c) Neighborhood power diversion<br>(d) Meter switching<br>(e) Hack meter to falsify reading |
| $s_5$ | $p_{t_i,n}$ =mean($\mathbf{P^*_n}$) | SM reports the average<br>of the readings over the day | (a) Hack meter to falsify reading |

Note: *start* and *end* are the **random** starting and ending time of energy theft/equipment faulty period.

University of Malaya

## 5.6 Summary of Chapter

This chapter reviews the data collection and test setup applied in this research study. The introduction starts off by discussing the smart metering data extracted from SM electricity trial database released by Sustainable Energy Authority of Ireland. In the next section, the hardware experimentation setup to collect real-world smart metering data was discussed. In the subsection of hardware experimentation, the software used to program the data logger and descriptions of the components in the test rig, were discussed in detail. Then, the test setup of the anomaly detection frameworks in Matlab was presented. In the last part of the chapter, the attack model which consists of a diverse set of NTL attack functions is generated and described such that it closely resemble the real-world energy fraud/meter irregularities scenarios in AMI.

In part to evaluate the performance and validate the reliability of both the LR-based and LP-based anomaly detection frameworks, simulation and hardware experimentation-based experiments were conducted. The hardware experimentation results obtained are correlated with Matlab simulation results in order to confirm the reliability of the proposed anomaly detection frameworks in real world. The frameworks validation results will be evaluated and discussed in detail in the next chapter.

University of Malaya

# CHAPTER 6: RESULTS AND DISCUSSIONS

## 6.1 Overview

In this chapter, the framework validation results are presented and evaluated. First of all, the metric used to evaluate the NTL detection frameworks is presented. Then, frameworks validation through data obtained from the Irish Smart Energy Trial Database released by Commission for Energy Regulation (2009) are described, whereby the regression and optimization results obtained from Matlab are discussed and evaluated in Section 6.3. Besides, more data are also extracted to study how the proposed frameworks scale with the number of consumers. Apart from scalability of the proposed frameworks, the impact caused by TLs and measurement noise on detection rate improvement is also investigated. In Section 6.4, frameworks validation through hardware experimentation is conducted. An AMI test rig is built in the laboratory to assess the performance and validate the reliability of the proposed frameworks in real SG environment. Next, a table is presented to *functionally* compare the proposed anomaly detection frameworks with existing NTL detection schemes, whereby the performance of the proposals are compared with the most recent and state-of-the-art energy theft detection schemes discussed in Section 2.4. Subsequently in Section 6.7, performance comparison studies are performed to study the strengths and weaknesses of the two proposed anomaly detection frameworks. Finally, a summary is given in Section 6.8 to conclude the proposed anomaly detection frameworks which were put forward in Chapters 3 and 4.

## 6.2 Performance Metric

For NTL detection, the goal is to increase the detection rate in order to discover as many NTL occurrences as possible, while reducing the number of false positives in order to minimize the number of costly onsite inspections. In order the evaluate the NTL detection

University of Malaya

model using a single performance measure, the detection rate $DR$, which is also known as *sensitivity* is computed as:

$$DR = \frac{TP}{TP + FN} \times 100\%, \tag{6.1}$$

whereby $TP$ denotes the number of true positives and $FN$ represents the number of false negatives.

## 6.3 Frameworks Validation Through Data from the Irish Smart Energy Trial

In this section, MLR and optimization analyses are presented to assess the performance of the proposed anomaly detection frameworks through data extracted from the Smart Meter Electricity Trial Database. According to Jokar et al. (2016) and Sahoo et al. (2015), real-world SG energy theft samples rarely, or do not, exist because SG is not fully implemented. As a result, the smart energy data from the Irish Smart Meter Electricity Trial denoted by $\mathbf{P}_n^d = \{p_{t_{1,n}}^d, p_{t_{2,n}}^d, \cdots, p_{t_{48,n}}^d\}$, for time $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_{48}\}$ on day $d \in \mathbf{D} = \{1, 2, \cdots, D\}$ for consumer $n \in \mathbf{N} = \{1, 2, \cdots, N\}$, are extracted from the Irish Smart Energy Trial (Commission for Energy Regulation, 2009) in this study. Then, different types of malicious scenario as discussed in Section 5.5 are generated based on the extracted benign trial dataset.

Two series of simulations in Matlab R2014b are conducted to evaluate the performance of the proposed LR-based and LP-based anomaly detection schemes. Specifically, two scenarios are considered, namely, fraudulent consumers steal at a *fixed rate* (constant anomaly coefficient) and *varying rate* (varying anomaly coefficient). Since the dataset released by Commission for Energy Regulation (2009) contains only consumers' energy consumption data, the collector measurement $c_{t_i}$ is obtained by duly summing up the energy consumption of all consumers in the service area at time interval $t_i$ when there are

University of Malaya

no energy thefts and defective SMs, viz., original untampered data. According to Nagi, Mohammad, Yap, Tiong, and Ahmed (2008); Nagi (2009), the NTLs faced by UPs in developing countries such as India, Pakistan, Bangladesh, Lebanon and Malaysia amount to an average of between 20% to 30%. Therefore, taking the worst case, it is assumed that approximately 30% of the consumers in the NAN are stealing energy and/or SMs are reporting more on their energy usage. In other words, 30% of the consumers and/or SMs in the NAN have a non-zero anomaly coefficient. In the simulations, service area of sizes 15 and 45 energy consumers are considered. The minimum time of anomaly is subject to the time granularity of the SM (i.e., one slot = 30 minutes).

### 6.3.1 Simulation for LR-based Detection Framework

As discussed previously in Section 3.4, both the LR-based anomaly detection schemes (i.e., LR-ETDM and CVLR-ETDM) do not consider TLs in the SGs. According to Sahoo et al. (2015), TLs can be computed by observing the data from DT and the current readings collected by conventional analog or smart power meters. Therefore, once the TLs are calculated, the proposed framework can be adjusted accordingly by subtracting TLs from vector $\mathbf{y}$ as expressed in Equation (3.2).

#### 6.3.1.1 Simulation: LR-ETDM

Here, it is assumed that the fraudulent consumers steal energy all the time and never stop cheating (i.e., case $s_2$ as discussed in the attack model, where $\nu \in (0, 0.95)$). At the same time, some of the SMs had malfunctioned throughout the period of observations (i.e., case $s_2$, where $\nu \in (1.05, 2.5]$). Therefore, the rates of cheating as well as reporting more (due to malfunctioning) *do not change* and hence the anomaly coefficients are *constant*.

Here, one-day half-hourly energy data (i.e., 48 data points) are extracted from the Irish Smart Energy Trial for the theft detection analyses. The constant cheating/malfunctioning

University of Malaya

**Table 6.1: Comparison between constant $a_n$ and $\tilde{a}_{n(LR)}$ obtained by LR-ETDM for the size of 15 consumers**

| Consumer $n$ | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(LR)}$ | $p$-value$_{\tilde{a}(LR)}$ | $\frac{1}{1+\tilde{a}_{n(LR)}}$ |
|---|---|---|---|---|---|---|
| 1 | Under-report by 50% | 1 | 0.50 | 1 | 1.6011e-180 | 0.50 |
| 4 | Over-report by 30% | −0.2308 | 1.30 | −0.2308 | 6.6226e-170 | 1.30 |
| 7 | Under-report by 50% | 1 | 0.50 | 1 | 2.7614e-176 | 0.50 |
| 11 | Over-report by 50% | −0.3333 | 1.50 | −0.3333 | 1.1096e-187 | 1.50 |
| 13 | Under-report by 60% | 1.5 | 0.40 | 1.5 | 4.5576e-185 | 0.40 |
| 15 | Under-report by 50% | 1 | 0.50 | 1 | 2.9711e-185 | 0.50 |
| Others | Honest | 0 | 1 | 0 | $> 0.01$ | 1 |

scenario for the size of 15 consumers is setup as shown in Table 6.1. The values of $a_n$ represent the exact state of each SM (i.e., honest, compromised or faulty) from the dataset. For instance, when consumer 1 under-reports what was consumed by 50% (i.e., consumes 1kWh but reports 0.5kWh, the meter discrepancy $y_{t_i} = 0.5$kWh), therefore $a_1 = \frac{y_{t_i}}{p_{t_i,1}} = \frac{0.5}{0.5} = 1$. As discussed in Section 4.4.2, the fraction of reported consumption of the consumer is computed as $\frac{1}{1+a_n} = \frac{p_{t_i,n}}{c_{t_i}}$. Thus, the fraction of reported consumption of consumer 1 is $\frac{p_{t_i,1}}{c_{t_i}} = \frac{0.5}{1.0} = 0.50$. Meanwhile, consumer 4 over-reports energy consumption by 30% (i.e., consumes 1kWh but reports 1.3kWh, the meter discrepancy $y_{t_i} = -0.3$kWh). In such a case, $a_4 = \frac{y_{t_i}}{p_{t_i,4}} = \frac{-0.3}{1.3} = -0.2308$ and the fraction of reported consumption of consumer 4 is $\frac{p_{t_i,4}}{c_{t_i}} = \frac{1.3}{1.0} = 1.30$. The other consumers who have $a_n = 0$ are honest because they report the actual energy consumption (i.e., consumes 0.5kWh and reports 0.5kWh, the meter discrepancy $y_{t_i}$ =0kWh). In such a case, $a_n = \frac{y_{t_i}}{p_{t_i,n}} = \frac{0}{0.5} = 0$ and the fraction of reported consumption of the consumer $n$ is $\frac{p_{t_i,n}}{c_{t_i}} = \frac{0.5}{0.5} = 1$.

The values of $\tilde{a}_{n(LR)}$ in Table 6.1 are the computed values obtained by LR-ETDM. Figure 6.1 depicts the computed values $\tilde{a}_{n(LR)}$. As shown in Figure 6.1, the proposed LR-ETDM scheme performs well for each of the cases, i.e., when there are 15 and 45 consumers in the service area. In particular, in the case of 15 consumers, it is observed from Table 6.1 that there are six consumers who have $p$-values less than 0.01, which

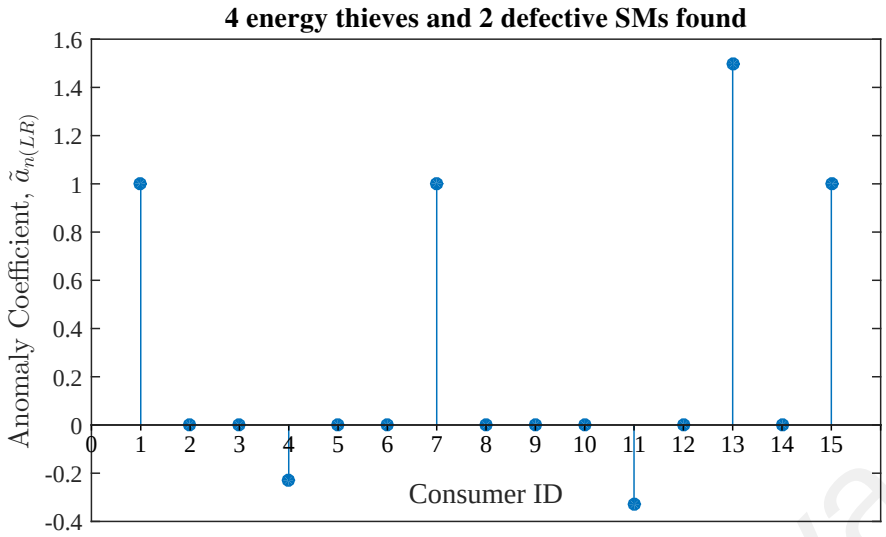
University of Malaya

implies that these consumers might have anomaly coefficients not equal to 0. Therefore, these suspected consumers are shortlisted for further investigation. It is obvious from Figure 6.1(a) that they have non-zero anomaly coefficients. Particularly, there are four energy thieves (i.e., consumers 1, 7, 13 and 15) who only report fraction of their energy consumption (i.e., $\tilde{a}_{n(LR)} > 0$). Meanwhile, two SMs (i.e., the 4-th and 11-th) are out of order as the meters report more than what the consumers actually consumed (i.e., $\tilde{a}_{n(LR)} < 0$). Furthermore, the nine honest consumers who have $p$-value$_{\tilde{a}(LR)} > 0.01$ and therefore $\tilde{a}_{n(LR)} = 0$ can also be identified easily. Based on these results, the data collector can effectively detect all the energy thieves as well as the malicious SMs, then computes how much less or more they have paid in their monthly bills by $\frac{1}{1+\tilde{a}_{n(LR)}}$. For example, the 11-th SM has $\tilde{a}_{11(LR)} = -0.3333$ and the fraction of reported consumption of consumer 11 is $\frac{1}{1+\tilde{a}_{11(LR)}} = \frac{1}{1+(-0.3333)} = 1.50$. Therefore, the 11-th SM is classified as malfunctioning for reporting 50% more than what was consumed. On a different note, consumer 13 has $\tilde{a}_{13(LR)} = 1.5$ and the fraction of reported usage of consumer 13 is $\frac{1}{1+\tilde{a}_{13(LR)}} = \frac{1}{1+1.5} = 0.40$. Thus, consumer 13 is classified as energy thief for only reporting 40% of what was consumed. Other consumers who have $p$-value$_{\tilde{a}(LR)} > 0.01$ are classified as honest because they have a zero anomaly coefficient (i.e., fraction of reported consumption $\frac{1}{1+\tilde{a}_{n(LR)}} = \frac{1}{1+0} = 1$).

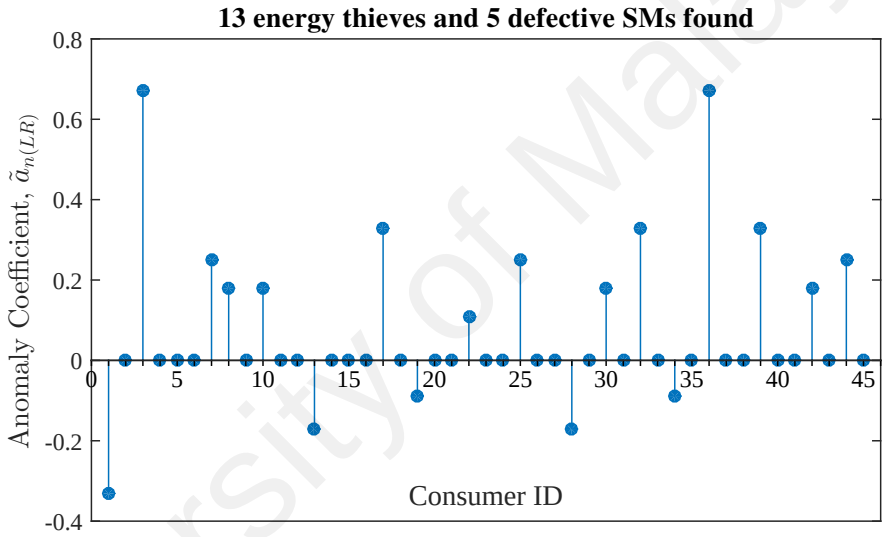
In order to study how MLR estimation scales with the number of consumers in the neighborhood, a NAN of 45 consumers is considered. Similar result is observed in Figure 6.1(b) for the case of 45 consumers, which is setup as shown in Table 6.2. By isolating the consumers who have $p$-values less than 1% significance level and anomaly coefficients not equal to 0, *all* the energy thieves and defective SMs in the NAN can be recognized effectively.

A detection rate of 100% was achieved by LR-ETDM for the experiment setup as

University of Malaya




(a) 15 consumers




(b) 45 consumers

**Figure 6.1: Value of $\tilde{a}_{n(LR)}$ obtained by LR-ETDM when $a_n$ is constant for the sizes of (a) 15 consumers and (b) 45 consumers.**

presented in both Tables 6.1 and 6.2 when $a_n$ is *constant* and TLs are non-existent. In other words, all fraudulent consumers/metering defects are correctly detected by LR-ETDM as anomalies from the total true anomalous consumers.

Besides, simulation is also conducted by using LR-ETDM when some fraudulent consumers are cheating inconsistently and some of them are stealing energy constantly which is setup as shown in Table 6.3. Specifically, some of the untruthful consumers are pilfering energy all the time and some of them are cheating on their energy consumption

University of Malaya

**Table 6.2: Comparison between constant $a_n$ and $\tilde{a}_{n(LR)}$ obtained by LR-ETDM for the size of 45 consumers**

| Consumer $n$ | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(LR)}$ | $p$-value$_{\tilde{a}(LR)}$ | $\frac{1}{1+\tilde{a}_{n(LR)}}$ |
|---|---|---|---|---|---|---|
| 1 | Over-report by 50% | −0.3333 | 1.50 | −0.3333 | 6.4345e-12 | 1.50 |
| 3 | Under-report by 40% | 0.6667 | 0.60 | 0.6667 | 3.7043e-13 | 0.60 |
| 7 | Under-report by 20% | 0.2500 | 0.80 | 0.2500 | 1.6617e-11 | 0.80 |
| 8 | Under-report by 15% | 0.1765 | 0.85 | 0.1765 | 3.4275e-11 | 0.85 |
| 10 | Under-report by 15% | 0.1765 | 0.85 | 0.1765 | 2.0045e-11 | 0.85 |
| 13 | Over-report by 20% | −0.1667 | 1.20 | −0.1667 | 8.8398e-12 | 1.20 |
| 17 | Under-report by 25% | 0.3333 | 0.75 | 0.3333 | 1.1502e-12 | 0.75 |
| 19 | Over-report by 10% | −0.0909 | 1.10 | −0.0909 | 1.9837e-11 | 1.10 |
| 22 | Under-report by 10% | 0.1111 | 0.90 | 0.1111 | 1.586e-11 | 0.90 |
| 25 | Under-report by 20% | 0.2500 | 0.80 | 0.2500 | 1.1088e-12 | 0.80 |
| 28 | Over-report by 20% | −0.1667 | 1.20 | −0.1667 | 7.6755e-12 | 1.20 |
| 30 | Under-report by 15% | 0.1765 | 0.85 | 0.1765 | 1.0358e-10 | 0.85 |
| 32 | Under-report by 25% | 0.3333 | 0.75 | 0.3333 | 1.0618e-14 | 0.75 |
| 34 | Over-report by 10% | −0.0909 | 1.10 | −0.0909 | 1.9761e-14 | 1.10 |
| 36 | Under-report by 40% | 0.6667 | 0.60 | 0.6667 | 1.7027e-15 | 0.60 |
| 39 | Under-report by 25% | 0.3333 | 0.75 | 0.3333 | 1.174e-14 | 0.75 |
| 42 | Under-report by 15% | 0.1765 | 0.85 | 0.1765 | 3.2694e-15 | 0.85 |
| 44 | Under-report by 20% | 0.2500 | 0.8 | 0.2500 | 5.5398e-15 | 0.80 |
| Others | Honest | 0 | 1 | 0 | $> 0.01$ | 1.00 |

only during a certain period in a day. The values of $\tilde{a}_{n(LR)}$ obtained by LR-ETDM are presented in Figure 6.2. As discussed in Chapter 3, LR-ETDM becomes unstable under this scenario. Based on the estimated anomaly coefficients and corresponding $p$-values, LR-ETDM finds five cheating consumers and a faulty SM only. But, in reality, there are five energy thieves and two faulty SMs. However, LR-ETDM accuses the honest consumer 10 and 14 wrongly as they have $p$-value$_{\tilde{a}(LR)} < 0.01$ and $\tilde{a}_{n(LR)} > 0$. Meanwhile, consumer 4, 7 and 13 are left unidentified because their $p$-value$_{\tilde{a}(LR)} > 0.01$ and $\tilde{a}_{n(LR)} = 0$.

The inaccuracies are due to the limiting factors of regression model. LR explicitly assumes that the anomaly coefficients $a_n$ do not change throughout the period of observation (Chambers & Dinsmore, 2014). In other words, LR presumes that if an energy thief cheats, he/she cheats at the same rate throughout the day. Thus, some of the fraudulent consumers could evade detection when they do not cheat all the time. To

University of Malaya

**Table 6.3: Comparison between varying $a_n$ and $\tilde{a}_{n(LR)}$ obtained by LR-ETDM for the size of 15 consumers**

| Consumer $n$ | Description | Affected Time Slot | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(LR)}$ | $p$-value$_{\tilde{a}(LR)}$ | $\frac{1}{1+\tilde{a}_{n(LR)}}$ |
|---|---|---|---|---|---|---|---|
| 1 | Under-report by 40% | All the time | 0.6667 | 0.60 | 1.3182 | 8.7257e-06 | 0.43 |
| 4 | Over-report by 30% | All the time | −0.2307 | 1.30 | *0 | 0.15968 | 1 |
| 7 | Under-report by 50% | On-peak (From $t_{16}$ to $t_{39}$) | 1 | 0.50 | *0 | 0.024895 | 1 |
| 8 | Under-report by 40% | All the time | 0.6667 | 0.60 | 0.7232 | 0.0081053 | 0.58 |
| 10 | Honest | All the time | 0 | 1 | *0.4660 | 0.00043485 | 0.68 |
| 11 | Over-report by 50% | On-peak (From $t_{16}$ to $t_{39}$) | −0.3333 | 1.5 | −0.3093 | 2.3554e-07 | 1.45 |
| 13 | Under-report by 60% | All the time | 1.50 | 0.4 | *0 | 0.41787 | 1 |
| 14 | Honest | All the time | 0 | 1 | *0.4507 | 0.00045457 | 0.69 |
| 15 | Under-report by 15% | All the time | 0.1765 | 0.85 | 0.4507 | 4.7724e-05 | 0.69 |
| Others | Honest | All the Time | 0 | 1 | 0 | $> 0.01$ | 1 |

* False Positive

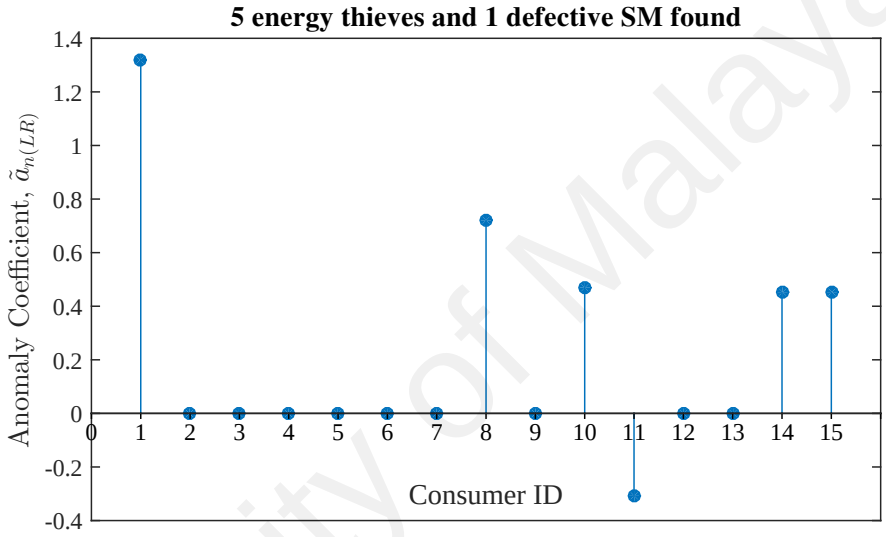



**Figure 6.2: Value of $\tilde{a}_{n(LR)}$ obtained by LR-ETDM when $a_n$ is varying (size of 15 consumers).**

overcome the deficiency of LR-ETDM, *categorical variables* are introduced in the LR model to accurately reveal the locations and periods (i.e., during off-peak, on-peak periods of a day or whole day) of energy theft or device failure. The simulation results for the varying cheating/malfunctioning scenarios will be discussed in the next section.

#### 6.3.1.2 Simulation: CVLR-ETDM

In this section, simulations are conducted for the situation when energy thieves under-report what was consumed and/or SMs over-report the energy consumption all the time/during a certain time (i.e., cases $s_2$, $s_3$ and $s_4$ in the attack model). In other words, the

University of Malaya

rates of under-reporting as well as over-reporting *change* and hence the anomaly coefficients are *varying*. The goal is to verify the viability of the proposed CVLR-ETDM in handling the constant/varying cheating and malfunctioning problems. In the simulations, each consumer commits energy theft during off-peak, on-peak or all the time. The consumers' power consumption data are observed over two days to increase the number of observations to mitigate the effect of over-fitting (Tetko et al., 1995).

The varying cheating/malfunctioning scenario for the size of 15 consumers is setup as shown in Table 6.4. Here, values of $a_n$ and $\beta_n$ are exact settings from the dataset. Recall that $a_n$ itself denotes the anomaly coefficients of consumers during off-peak period while $(a_n + \beta_n)$ denotes the anomaly coefficients of consumers during on-peak hours. For instance, consumer 1 under-reports what was consumed by 20% only during *off-peak* period (i.e., consumes 1kWh but reports 0.8kWh, meter discrepancy $y_{t_i} = 0.2$kWh). Therefore, $a_1 = \frac{y_{t_i}}{p_{t_i,1}} = \frac{0.2}{0.8} = 0.25$ and $(a_1 + \beta_1) = 0$. On the other hand, consumer 3 under-reports what was consumed by 30% *all the time* (i.e., consumes 1kWh but reports 0.7kWh, meter discrepancy $y_{t_i} = 0.3$kWh). Hence, $a_3 = \frac{y_{t_i}}{p_{t_i,3}} = \frac{0.3}{0.7} = 0.4286$ and $(a_3 + \beta_3) = 0.4286$. Meanwhile, consumer 6 under-reports energy consumption by 25% only during *on-peak* period (i.e., consumes 1kWh but reports 0.75kWh, meter discrepancy $y_{t_i} = 0.25$kWh). In such a case, $a_6 = 0$ and $(a_6 + \beta_6) = \frac{y_{t_i}}{p_{t_i,6}} = \frac{0.25}{0.75} = 0.3333$.

The values of $\tilde{a}_{n(CVLR)}$ and $\tilde{\beta}_{n(CVLR)}$ in Table 6.4 are the computed coefficients obtained by CVLR-ETDM. Figure 6.3 depicts the computed values $\tilde{a}_{n(CVLR)}$ and $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$. In the figure, black bar represents off-peak period $\tilde{a}_{n(CVLR)}$ (i.e., varying anomaly coefficient) and white bar represents on-peak period $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ (i.e., varying anomaly coefficient). If black bar and white bar co-appear (i.e., constant anomaly coefficient), it implies that the energy frauds occur or defective meters exist all the time. Results in Figure 6.3(a) suggest that there are five dishonest consumers and a faulty SM in

University of Malaya

the service area of 15 consumers. It can be seen from Table 6.4 that $p$-values of $\tilde{\beta}_{1(CVLR)}$, $\tilde{\beta}_{6(CVLR)}$, $\tilde{\beta}_{12(CVLR)}$ and $\tilde{\beta}_{15(CVLR)}$ are less than a 1% significance level, indicating that consumers 1, 6, 12 and 15 are likely to have $\tilde{\beta}_{n(CVLR)} \neq 0$. Recall that a non-zero $\tilde{\beta}_{n(CVLR)}$ implies that consumer $n$ has different cheating patterns throughout the period of observations (i.e., varying anomaly coefficient). In particular, consumer 1 and consumer 15 steal energy (i.e., $\tilde{a}_{n(CVLR)} > 0$ and $\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)} = 0$, where $\tilde{\beta}_{n(CVLR)} = -\tilde{a}_{n(CVLR)}$) only during *off-peak* period (i.e., black bar appears) while consumer 6 and consumer 12 pilfer energy (i.e., $\tilde{a}_{n(CVLR)} = 0$ and $\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)} > 0$) only during *on-peak* period (i.e., white bar appears). On the other hand, it is observed that $p$-values of both $\tilde{\beta}_{3(CVLR)}$ and $\tilde{\beta}_{9(CVLR)}$ are greater than 0.01, indicating that both consumer 3 and consumer 9 have $\tilde{\beta}_{n(CVLR)} = 0$ and hence they do not change their cheating behaviors. In other words, either the consumer is stealing or the SM is defective all the time when his/her $\tilde{a}_{n(CVLR)}$ is non-zero. Specifically, consumer 3 is stealing energy all the time (i.e., black and white bars co-appear, $\tilde{a}_{n(CVLR)} > 0$, $\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)} > 0$) on both off-peak and on-peak periods. The 9-th SM is out of order all the time (i.e., black and white bars co-appear, $\tilde{a}_{n(CVLR)} < 0$, $\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)} < 0$).

Generally, data collector is placed on the DS at a utility selected interval, such as every 50 consumers per phase. In order to study the scalability of CVLR-ETDM, a NAN of 45 consumers is considered. Similar result is obtained for the case of 45 consumers, where the varying cheating/malfunctioning scenario is setup as shown in Table 6.5. The results are presented in Figure 6.3(b). Based on these findings, the data collector can calculate how much less/more the consumers have paid by analyzing the value of the anomaly coefficients and detection coefficients of the consumers as discussed in Section 4.4.2. However, the computation for fraction of reported consumption of each consumer is omitted from Tables 6.4 and 6.5 due to space constraints.

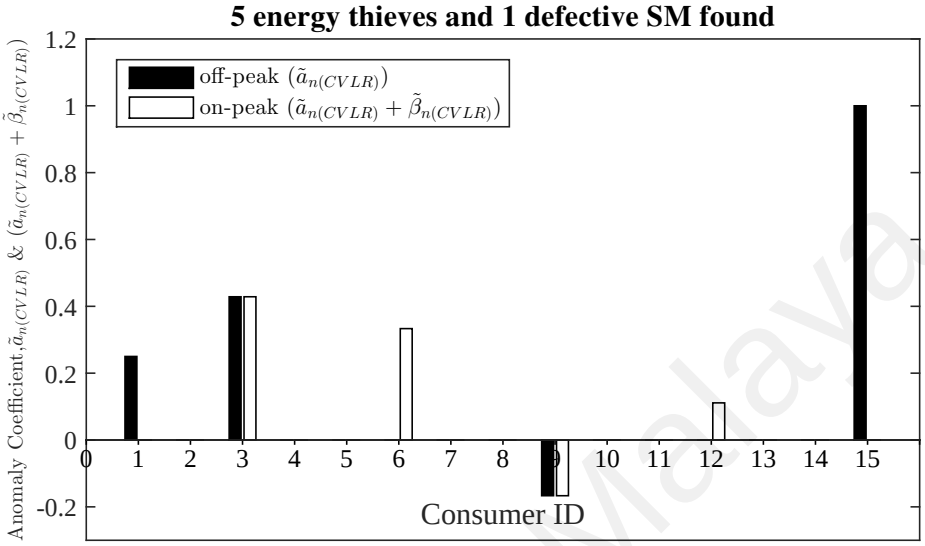
University of Malaya

**Table 6.4: Comparison between $a_n$ & $\beta_n$ and $\tilde{a}_{n(CVLR)}$ & $\tilde{\beta}_{n(CVLR)}$ obtained by CVLR-ETDM for the size of 15 consumers**

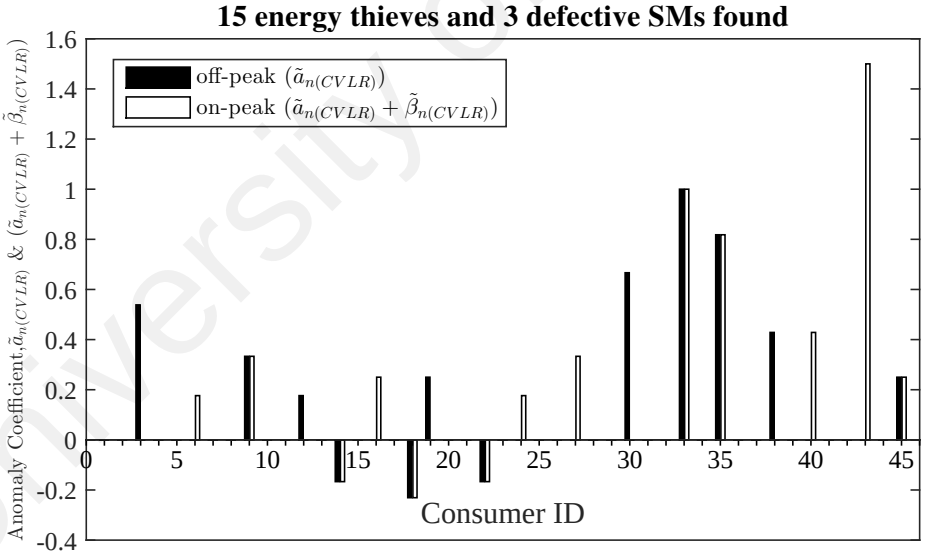
| Consumer $n$ | Description | Affected Time Slot | $a_n$ | $\beta_n$ | $(a_n + \beta_n)$ | $\tilde{a}_{n(CVLR)}$ | $p$-value$_{\tilde{a}(CVLR)}$ | $\tilde{\beta}_{n(CVLR)}$ | $p$-value$_{\tilde{\beta}(CVLR)}$ | $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Under-report by 20% | Off-peak | 0.2500 | −0.2500 | 0 | 0.2500 | 0 | −0.2500 | 0 | 0 |
| 3 | Under-report by 30% | All the time | 0.4286 | 0 | 0.4286 | 0.4286 | 0 | $-7.96E-14 \approx 0$ | 1 | 0.4286 |
| 6 | Under-report by 25% | On-peak | 0 | 0.3333 | 0.3333 | 0 | 1 | 0.3333 | 0 | 0.3333 |
| 9 | Over-report by 20% | All the time | −0.1667 | 0 | −0.1667 | −0.1667 | 0 | $-3.74E-14 \approx 0$ | 1 | −0.1667 |
| 12 | Under-report by 10% | On-peak | 0 | 0.1111 | 0.1111 | 0 | 1 | 0.1111 | 1.751e-320 | 0.1111 |
| 15 | Under-report by 50% | Off-peak | 1 | −1 | 0 | 1 | 0 | −1 | 0 | 0 |
| Others | Honest | All the time | 0 | 0 | 0 | 0 | $> 0.01$ | 0 | $> 0.01$ | 0 |

Note: Off-peak (From $t_1$ to $t_{15}$ and from $t_{40}$ to $t_{48}$), On-peak (From $t_{16}$ to $t_{39}$)

University of Malaya




(a) 15 consumers




(b) 45 consumers

**Figure 6.3: Value of $\tilde{a}_{n(CVLR)}$ and ($\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)}$) obtained by CVLR-ETDM when $a$ is varying for the sizes of (a) 15 consumers and (b) 45 consumers.**

University of Malaya

**Table 6.5: Comparison between $a_n$ & $\beta_n$ and $\tilde{a}_{n(CVLR)}$ & $\tilde{\beta}_{n(CVLR)}$ obtained by CVLR-ETDM for the size of 45 consumers**

| Consumer $n$ | Description | Affected Time Slot | $a_n$ | $\beta_n$ | $(a_n + \beta_n)$ | $\tilde{a}_{n(CVLR)}$ | $p$-value$_{\tilde{a}(CVLR)}$ | $\tilde{\beta}_{n(CVLR)}$ | $p$-value$_{\tilde{\beta}(CVLR)}$ | $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | Under-report by 35% | Off-peak | 0.5385 | −0.5385 | 0 | 0.5385 | 2.25E-17 | −0.5385 | 2.38E-17 | 0 |
| 6 | Under-report by 15% | On-peak | 0 | 0.1765 | 0.1765 | 0 | 1 | 0.1765 | 1.56E-13 | 0.1765 |
| 9 | Under-report by 25% | All the time | 0.3333 | 0 | 0.3333 | 0.3333 | 1.67E-12 | 0 | 1 | 0.3333 |
| 12 | Under-report by 15% | Off-peak | 0.1765 | −0.1765 | 0 | 0.1765 | 3.37E-14 | −0.1765 | 3.69E-14 | 0 |
| 14 | Over-report by 20% | All the time | −0.1667 | 0 | −0.1667 | −0.1667 | 9.98E-15 | 0 | 1 | −0.1667 |
| 16 | Under-report by 20% | On-peak | 0 | 0.2500 | 0.2500 | 0 | 1 | 0.2500 | 1.95E-13 | 0.2500 |
| 18 | Over-report by 30% | All the time | −0.2308 | 0 | −0.2308 | −0.2308 | 1.85E-16 | 0 | 1 | −0.2308 |
| 19 | Under-report by 20% | Off-peak | 0.2500 | −0.2500 | 0 | 0.2500 | 4.60E-16 | −0.2500 | 4.81E-16 | 0 |
| 22 | Over-report by 20% | All the time | −0.1667 | 0 | −0.1667 | −0.1667 | 3.64E-20 | 0 | 1 | −0.1667 |
| 24 | Under-report by 15% | On-peak | 0 | 0.1765 | 0.1765 | 0 | 1 | 0.1765 | 1.65E-17 | 0.1765 |
| 27 | Under-report by 25% | On-peak | 0 | 0.3333 | 0.3333 | 0 | 1 | 0.3333 | 4.56E-22 | 0.3333 |
| 30 | Under-report by 40% | Off-peak | 0.6667 | −0.6667 | 0 | 0.6667 | 2.27E-18 | −0.6667 | 2.29E-18 | 0 |
| 33 | Under-report by 50% | All the time | 1 | 0 | 1 | 1 | 6.35E-26 | 0 | 1 | 1 |
| 35 | Under-report by 45% | All the time | 0.8182 | 0 | 0.8182 | 0.8182 | 3.52E-25 | 0 | 1 | 0.8182 |
| 38 | Under-report by 30% | Off-peak | 0.4286 | −0.4286 | 0 | 0.4286 | 8.56E-20 | −0.4286 | 8.66E-20 | 0 |
| 40 | Under-report by 30% | On-peak | 0 | 0.4286 | 0.4286 | 0 | 1 | 0.4286 | 3.74E-22 | 0.4286 |
| 43 | Under-report by 60% | On-peak | 0 | 1.5000 | 1.5000 | 0 | 1 | 1.5000 | 2.18E-25 | 1.5000 |
| 45 | Under-report by 20% | All the time | 0.2500 | 0 | 0.2500 | 0.2500 | 1.59E-23 | 0 | 1 | 0.2500 |
| Others | Honest | All the time | 0 | 0 | 0 | 0 | $> 0.01$ | 0 | $> 0.01$ | 0 |

Note: Off-peak (From $t_1$ to $t_{15}$ and from $t_{40}$ to $t_{48}$), On-peak (From $t_{16}$ to $t_{39}$)

Universiti Malaya

Meanwhile, when a fraudulent consumer $n$ attempts to send zero readings all the time or during a certain period in the day, the $p$-value of the $a_n$ will show not a number (NaN) in the Matlab regression analysis. In such a case, the SM of the dishonest consumer $n$ should be inspected and replaced before the proposed LR-based schemes are re-invoked to obtain a more accurate regression analysis.

Similarly, a detection rate of 100% was achieved by CVLR-ETDM for the experiment setup as shown in Tables 6.4 and 6.5 when $a_n$ is *varying* and TLs are non-existent i.e., all the dishonest consumers/metering defects are identified accurately by CVLR-ETDM as anomalies from the total true anomalous consumers.

### 6.3.2 Simulation for LP-based Detection Framework

In the LR-based anomaly detection framework, TLs in the SG are not considered. Specifically, the LR-based framework assumes that the power line losses are known, which in practice may be difficult to acquire. In the pursuit of higher anomaly detection rate and lower false positives, the assumption of known power line losses is relaxed and a new LP-based anomaly detection framework that can overcome the deficiency of the LR-based anomaly detection framework is proposed. Particularly, the impact caused by TLs and measurement noise/error on the detection analysis is taken into account in the design of anomaly detection framework. Recall that, *loss factor* and *error term* are introduced for capturing the percentage of TLs and amount of measurement noise, respectively, in the service area.

According to Au et al. (2008), the average TLs of LV network in Malaysia was reported to range from 0.59% to 3.23%, subject to percentage loading of the LV network. Besides, according to Accenture (2011), normal line losses should be in the 0.5 to 4 percent range when there is no energy diversion. To show the viability of the proposed framework in estimating the amount of TLs, an evaluation environment with $3\% - 5\%$ of TLs is created,

University of Malaya

i.e.,

$$c_{t_i}^{d} \leftarrow \left( \sum_{n=1}^{N} p_{t_i,n}^{d*} \right) \div \mu, \mu \in [0.95, 0.97], \forall t_i \in \mathbf{T}, \forall d \in \mathbf{D}, \tag{6.2}$$

whereby randomized $\mu$ are generated for each time period $t_i$ which lead to different injected TLs values.

Note that state-of-the-art SMs record very accurate measurements, where the errors are usually modeled by white uncorrelated noise with zero mean and standard deviation (S. A. Salinas & Li, 2015). Therefore, to further validate the proposed framework, the errors of measurement are included as below:

$$y_{t_i}^{d} = c_{t_i}^{d} - \sum_{n=1}^{N} p_{t_i,n}^{d} + e_{t_i}^{d}, \tag{6.3}$$

where $e_{t_i}^{d}$ is the white uncorrelated noise with zero mean and standard deviation of 0.01 at time interval $t_i$ on day $d$.

#### 6.3.2.1 Simulation: ADF

In this section, the performance of ADF when energy thieves steal electricity at a constant rate continuously (i.e., $s_2 \in (0, 0.95)$) and/or SMs are out of order all the time (i.e., $s_2 \in (1.05, 2.5]$) is evaluated. The constant cheating/malfunctioning scenario for the size of 15 consumers is setup as shown in Table 6.6. Here, one-day half-hourly energy data (i.e., 48 data points) are extracted from the Irish Smart Energy Trial for the theft detection analysis. The values of $a_n$ represent the exact state of each SM (i.e., honest, compromised or defective) from the dataset, whereas $\tilde{a}_{n(ADF)}$ denotes the computed anomaly coefficients obtained by ADF.

Figure 6.4 depicts the computed $\tilde{a}_{n(ADF)}$ for the service area consisting of 15 consumers.

University of Malaya

**Table 6.6: Comparison between constant $a_n$ and $\tilde{a}_{n(ADF)}$ obtained by ADF for the size of 15 consumers**

| Consumer $n$ | Type of Consumer | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(ADF)}$ | $\frac{1}{1+\tilde{a}_{n(ADF)}}$ |
|---|---|---|---|---|---|---|
| 1 | Residential | Under-report by 40% | 0.67 | 0.60 | 0.8315 | 0.55 |
| 2 | Commercial | Honest | 0 | 1 | $0.01 \approx 0$ | 1 |
| 3 | Residential | Honest | 0 | 1 | $0.03 \approx 0$ | 1 |
| 4 | Commercial | Over-report by 30% | −0.23 | 1.30 | −0.23 | 1.30 |
| 5 | Commercial | Honest | 0 | 1 | $-0.02 \approx 0$ | 1 |
| 6 | Residential | Honest | 0 | 1 | $0.01 \approx 0$ | 1 |
| 7 | Commercial | Under-report by 50% | 1 | 0.50 | 1.02 | 0.50 |
| 8 | Commercial | Honest | 0 | 1 | 0 | 1 |
| 9 | Residential | Honest | 0 | 1 | $0.01 \approx 0$ | 1 |
| 10 | Commercial | Honest | 0 | 1 | $-0.03 \approx 0$ | 1 |
| 11 | Residential | Honest | 0 | 1 | 0 | 1 |
| 12 | Commercial | Honest | 0 | 1 | 0 | 1 |
| 13 | Residential | Over-report by 50% | −0.33 | 1.50 | −0.39 | 1.64 |
| 14 | Residential | Honest | 0 | 1 | $0.01 \approx 0$ | 1 |
| 15 | Commercial | Under-report by 60% | 1.50 | 0.40 | 1.37 | 0.42 |

Results of $\tilde{a}_{n(ADF)}$ in Table 6.6 and Figure 6.4 suggest that there are five anomalous consumers with $\tilde{a}_{n(ADF)}$ not in $[-0.05, 0.05]$. Recall that, consumers who have anomaly coefficients in $[-0.05, 0.05]$ are assumed to be honest (i.e., $\tilde{a}_{n(ADF)} \approx 0$). Specifically, consumers 1, 7 and 15 pilfer energy all the time (since $\tilde{a}_{n(ADF)} > 0.05$) while the 4-th and 13-th SMs are faulty (since $\tilde{a}_{n(ADF)} < -0.05$). Subsequently, the collector can compute the fraction of reported consumption by $\frac{1}{1+\tilde{a}_{n(ADF)}}$ based on the computed anomaly coefficients as shown in Table 6.6. For instance, it can be seen that consumer 1 reports approximately 55% of what was consumed since $\frac{1}{1+0.8315} \approx 0.55$, while the 4-th SM over-reports what was consumed by 30% since $\frac{1}{1+(-0.23)} \approx 1.3$. Meanwhile, the second consumer has $\frac{1}{1+0} = 1$, which classifies him as honest. Note that, the slight differences between $a_n$ and the computed $\tilde{a}_{n(ADF)}$ in Table 6.6 are likely due to the injected measurement noise in Equation (6.3).

Meanwhile, Figure 6.5 depicts the values of the actual amount of TLs, $l_{t_i}$ and the computed loss factors, $\tilde{l}_{t_i}$ obtained by ADF over 48 time intervals in a day. The results in Figure 6.5 suggest that ADF is able to predict the percentage of TLs of each time interval

University of Malaya

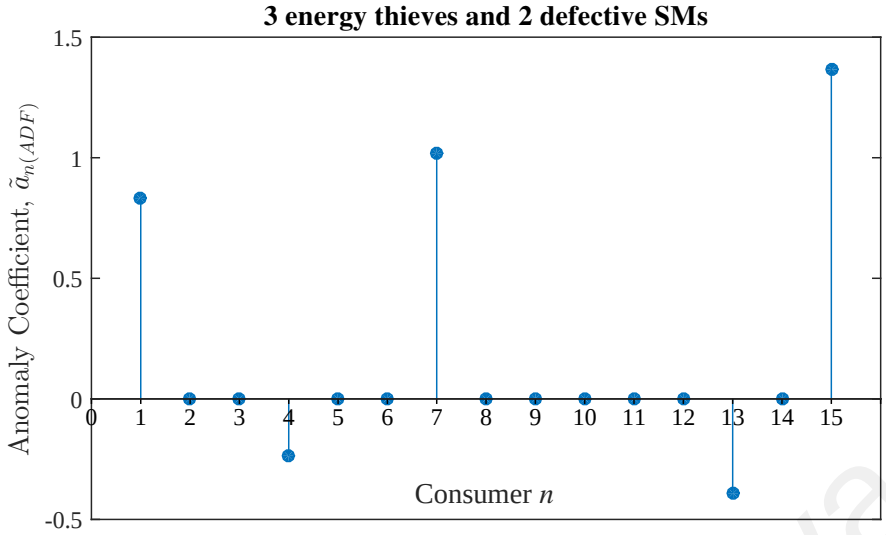



**Figure 6.4: Value of anomaly coefficients, $\tilde{a}_{n(ADF)}$ obtained by ADF when $a_n$ is constant (size of 15 consumers).**

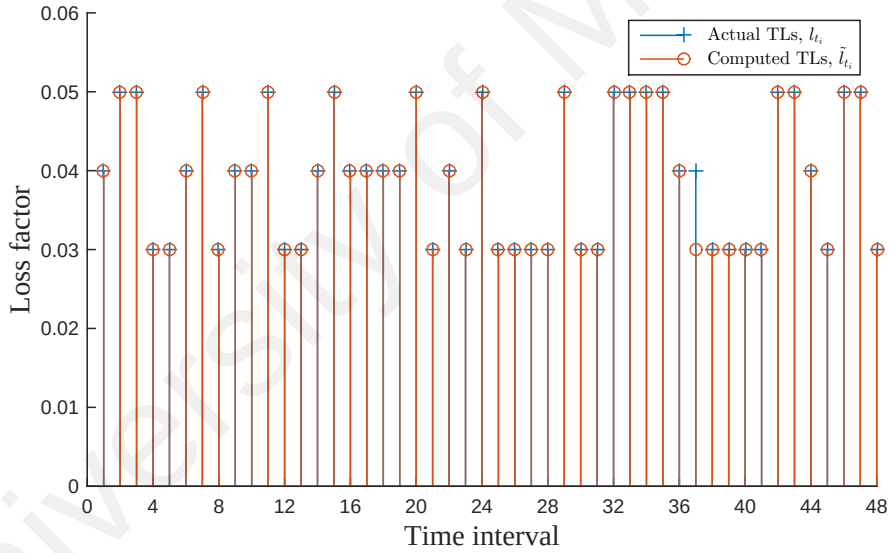



**Figure 6.5: Value of loss factors, $\tilde{l}_{t_i}$ obtained by ADF over 48 time intervals (size of 15 consumers).**

up to 98% accuracy (i.e., except the 37-th time slot, with a deviation of 1%) even the amount of TLs are injected randomly as shown in Equation (6.2).

In order to study how the proposed framework scales with the number of consumers, a service area of 45 consumers is considered. The simulation for the scenario of 45 consumers can be setup in a similar manner, and the detailed descriptions are presented

University of Malaya

**Table 6.7: Comparison between constant $a_n$ and $\tilde{a}_{n(ADF)}$ obtained by ADF for the size of 45 consumers**

| Consumer $n$ | Type of Consumer | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(ADF)}$ | $\frac{1}{1+\tilde{a}_{n(ADF)}}$ |
|---|---|---|---|---|---|---|
| 1 | Commercial | Over-report by 90% | −0.47 | 1.9 | -0.48 | 1.92 |
| 8 | Residential | Under-report by 50% | 1.00 | 0.50 | 0.99 | 0.50 |
| 10 | Commercial | Under-report by 40% | 0.67 | 0.60 | 0.66 | 0.60 |
| 19 | Commercial | Under-report by 50% | 1.00 | 0.50 | 1.00 | 0.50 |
| 22 | Commercial | Under-report by 70% | 2.33 | 0.30 | 2.35 | 0.30 |
| 27 | Commercial | Under-report by 40% | 0.67 | 0.60 | 0.67 | 0.60 |
| 32 | Commercial | Under-report by 50% | 1.00 | 0.50 | 1.00 | 0.50 |
| 34 | Residential | Under-report by 30% | 0.43 | 0.70 | 0.42 | 0.70 |
| 38 | Residential | Under-report by 40% | 0.67 | 0.60 | 0.66 | 0.60 |
| 42 | Commercial | Under-report by 60% | 1.50 | 0.40 | 1.51 | 0.40 |
| Others | - | Honest | 0 | 1 | 0 | 1 |

in Table 6.7. Four-day half-hourly energy consumption data (i.e., 192 data points) are extracted from the Irish Smart Energy Trial for the detection analysis. The metered consumption data are observed over a longer period for service area consisting of more consumers so that the proposed framework produces more accurate detection and avoids false positives. This is because observation of metered data over longer periods leads to addition in the number of constraints that can improve the accuracy of the NTL detection analysis. Figures 6.6 and 6.7 show corresponding results for the case of 45 consumers. The collector can detect all the faulty and/or compromised SMs and estimate the percentages of TLs correctly based on these results. Results from both service areas of 15 and 45 consumers suggest that ADF can detect the localities of energy fraud and metering defects regardless of the type of consumer and contracted power.

The results suggest that a detection rate of 100% was achieved by ADF for the constant cheating/malfunctioning scenarios as shown in both Tables 6.6 and 6.7 even in the presence of TLs and measurement noise.

University of Malaya

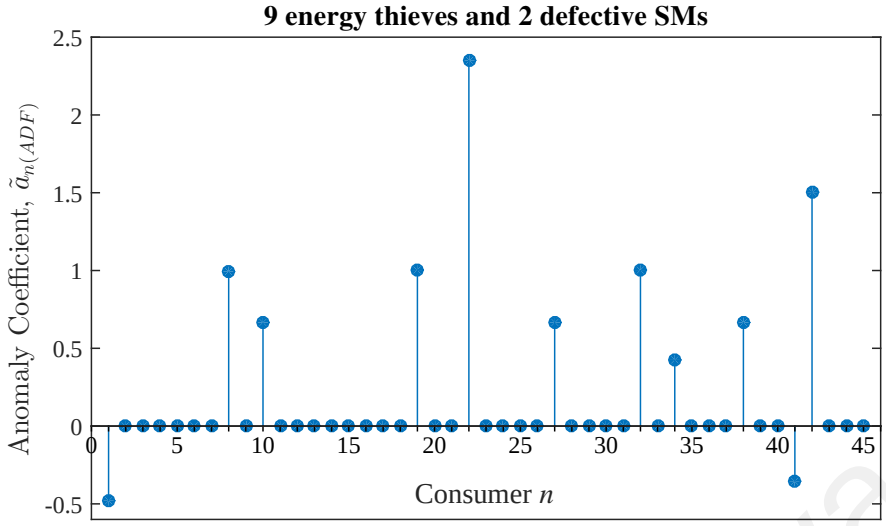



**Figure 6.6: Value of anomaly coefficients, $\tilde{a}_{n(ADF)}$ obtained by ADF when $a_n$ is constant (size of 45 consumers).**

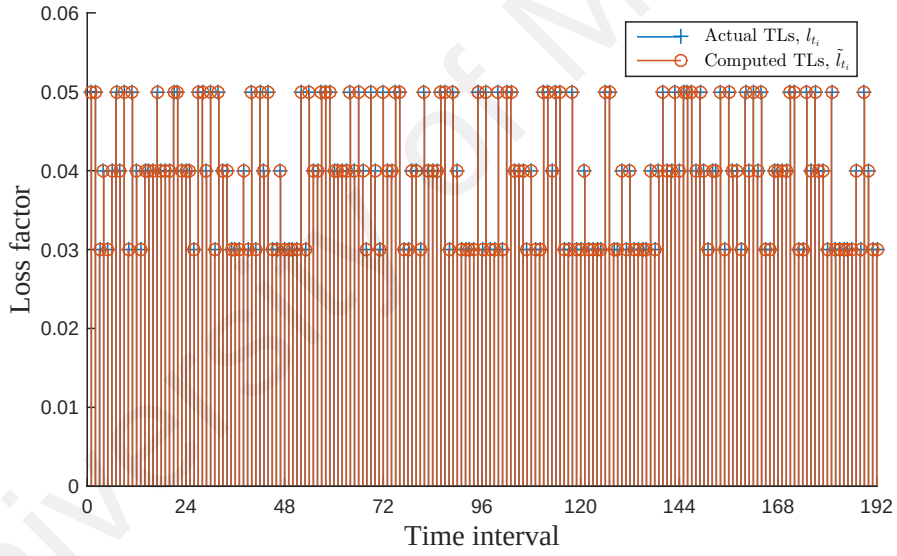



**Figure 6.7: Value of loss factors, $\tilde{l}_{t_i}$ obtained by ADF over 192 time intervals (size of 45 consumers).**

#### 6.3.2.2 Simulation: Enhanced ADF

In this section, the case where there are energy fraudsters who only cheat on their energy reporting during an intermittent period of the day is considered. The varying cheating/malfunctioning scenario for the size of 15 consumers is setup as shown in Table 6.8. Thirty-day energy consumption data (i.e., $D = 30$) are extracted for solving the

University of Malaya

varying cheating problems. Here, the values of $a_{t_i,n}$ represent the exact state of each SM (i.e., honest, compromised or defective) from the dataset at each time interval $t_i$. In the simulation setup, consumer 1 under-reports what was consumed by 60% (i.e., consumes 1 kWh but reports 0.4kWh, the meter discrepancy $y^d_{t_i} = 0.6$ kWh) only during time interval $t_i = \{t_{20}, t_{21}, \cdots, t_{37}\}$. Therefore, $a_{t_i,1} = \frac{y^d_{t_i}}{p^d_{t_i,n}} = \frac{0.6}{0.4} = 1.50$ while fraction of reported consumption of consumer 1 is $\frac{p^d_{t_i,1}}{c^d_{t_i}} = \frac{0.4}{1.0} = 0.40$ during the affected time slots where $t_i = \{t_{20}, t_{21}, \cdots, t_{37}\}$. On the other hand, consumer 15 over-reports what was consumed by 50% (i.e., consumes 1 kWh but reports 1.5kWh, the meter discrepancy $y^d_{t_i} = -0.5$ kWh) all the time. In such a case, $a_{t_1,15} = \frac{y^d_{t_i}}{p^d_{t_i,n}} = \frac{-0.5}{1.5} = -0.33$ while fraction of reported consumption of consumer 15 is $\frac{p^d_{t_i,15}}{c^d_{t_i}} = \frac{1.5}{1.0} = 1.50$ all the time where $t_i = \{t_1, t_2, \cdots, t_{48}\}$.

The values of $\overline{\tilde{a}}_{t_i,n(EADF)}$ in Table 6.8 are the average values of the computed $\tilde{a}_{t_i,n(EADF)}$ during the affected time slots, whenever $\tilde{a}_{t_i,n(EADF)} \notin [-0.05, 0.05]$. Again, recall that consumers who have anomaly coefficients in $[-0.05, 0.05]$ are assumed to be honest (i.e., $\tilde{a}_{t_i,n(EADF)} \approx 0$). Figure 6.8 depicts the values of $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced ADF for the size of 15 consumers. It is observed that there are four dishonest consumers and a faulty SM. Specifically, consumers 4 and 11 compromise SMs to steal energy all the time (since $\tilde{a}_{t_i,n(EADF)} > 0.05, \forall t_i \in \mathbf{T}$) while the 15-th SM is faulty all the time (since $\tilde{a}_{t_i,15(EADF)} < -0.05, \forall t_i \in \mathbf{T}$). Meanwhile, consumer 1 tampers SM to under-report energy consumption from the 20-th to 37-th time slots (since $\tilde{a}_{t_{20},1(EADF)}, \tilde{a}_{t_{21},1(EADF)}, \cdots, \tilde{a}_{t_{37},1(EADF)} > 0.05$), and consumer 10 only steals electricity from the 18-th to 36-th time slots (since $\tilde{a}_{t_{18},10(EADF)}, \tilde{a}_{t_{19},10(EADF)}, \cdots, \tilde{a}_{t_{36},10(EADF)} > 0.05$) of the day. The other consumers are classified as honest as their $\overline{\tilde{a}}_{t_i,n(EADF)}$ are approximately 0. Again, the slight differences between $a_{t_i,n}$ and the computed $\overline{\tilde{a}}_{t_i,n(ADF)}$ in Table 6.8 are caused by the injected measurement noise in Equation (6.3). As discussed earlier, the collector can also compute

University of Malaya

**Table 6.8: Comparison between varying $a_{t_i,n}$ and $\bar{\tilde{a}}_{t_i,n(EADF)}$ obtained by Enhanced ADF for the size of 15 consumers**

| Consumer $n$ | Type of Consumer | Description | Affected Time Slot, $t_i$ | $a_{t_i,n}$ | $\frac{1}{1+a_{t_i,n}}$ | $\bar{\tilde{a}}_{t_i,n(EADF)}$ | $\frac{1}{1+\bar{\tilde{a}}_{t_i,n(EADF)}}$ |
|---|---|---|---|---|---|---|---|
| 1 | Commercial | Under-report by 60% | From $t_{20}$ to $t_{37}$ | 1.50 | 0.40 | 1.45 | 0.41 |
| 4 | Commercial | Under-report by 50% | All the time | 1 | 0.50 | 0.97 | 0.51 |
| 10 | Commercial | Under-report by 50% | From $t_{18}$ to $t_{36}$ | 1 | 0.50 | 1.02 | 0.50 |
| 11 | Residential | Under-report by 40% | All the time | 0.67 | 0.60 | 0.65 | 0.61 |
| 15 | Residential | Over-report by 50% | All the time | −0.33 | 1.50 | −0.32 | 1.47 |
| Others | - | Honest | All the time | 0 | 1 | 0 | 1 |

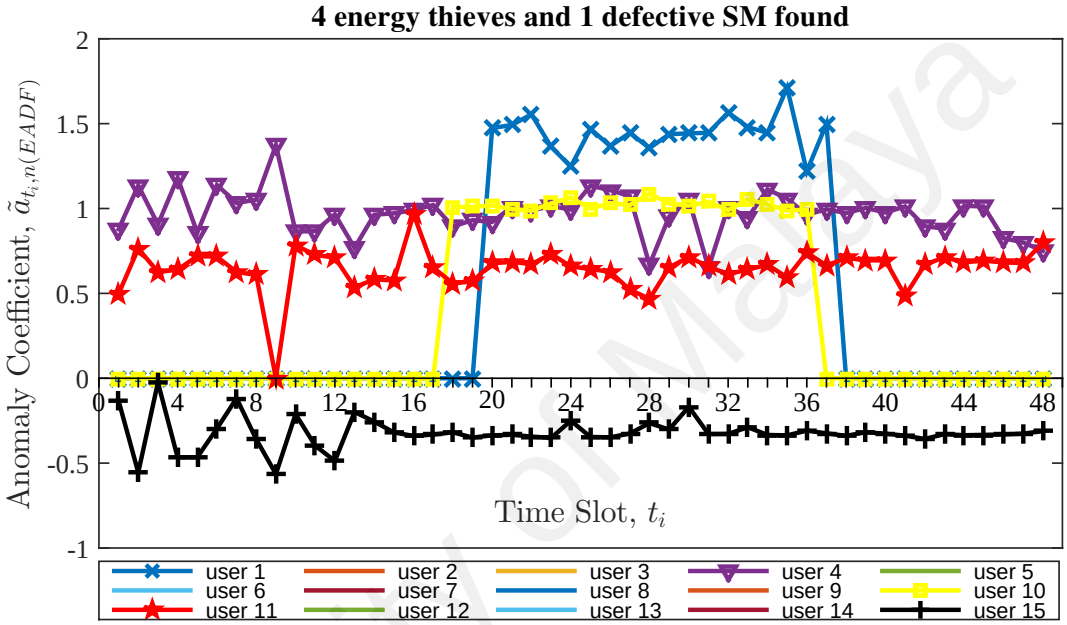



**Figure 6.8: Value of anomaly coefficients, $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced ADF when $a_{t_i,n}$ is varying (size of 15 consumers). Only anomalous cases are plotted.**

the fraction of reported energy usage based on the computed $\bar{\tilde{a}}_{t_i,n(EADF)}$ during the affected time slots, as shown in Table 6.8.

To evaluate the scalability of Enhanced ADF, a NAN of 45 consumers is considered. The simulation for the case of 45 consumers is setup in a similar manner, and the detailed descriptions are presented in Table 6.9. Similar results are observed in Table 6.9 and Figure 6.9. Here, 150-day energy consumption data (i.e., $D = 150$) are observed for solving the varying cheating problems in a larger service area. The results suggest that the proposed Enhanced ADF is able to detect energy thieves even if the anomalous consumers attempt to steal energy at intermittent periods. For instance, it can be seen from Figure 6.9

University of Malaya

**Table 6.9: Comparison between varying $a_{t_i,n}$ and $\overline{\tilde{a}}_{t_i,n(EADF)}$ obtained by Enhanced ADF for the size of 45 consumers**

| Consumer $n$ | Type of Consumer | Description | Affected Time Slot, $t_i$ | $a_{t_i,n}$ | $\frac{1}{1+a_{t_i,n}}$ | $\overline{\tilde{a}}_{t_i,n(EADF)}$ | $\frac{1}{1+\overline{\tilde{a}}_{t_i,n(EADF)}}$ |
|---|---|---|---|---|---|---|---|
| 1 | Commercial | Under-report by 50% | From $t_{19}$ to $t_{34}$ | 1 | 0.50 | 1.13 | 0.47 |
| 7 | Residential | Under-report by 70% | From $t_{21}$ to $t_{31}$ | 2.33 | 0.30 | 2.27 | 0.31 |
| 9 | Commercial | Under-report by 70% | From $t_{19}$ to $t_{37}$ | 2.33 | 0.30 | 2.28 | 0.30 |
| 10 | Commercial | Under-report by 50% | All the time | 1 | 0.50 | 1.03 | 0.49 |
| 19 | Commercial | Under-report by 60% | All the time | 1.50 | 0.40 | 1.52 | 0.40 |
| 22 | Commercial | Under-report by 40% | From $t_{17}$ to $t_{40}$ | 0.67 | 0.60 | 0.67 | 0.60 |
| 23 | Commercial | Under-report by 40% | From $t_4$ to $t_{21}$ | 0.67 | 0.60 | 0.72 | 0.58 |
| 25 | Commercial | Over-report by 80% | All the time | −0.44 | 1.80 | −0.29 | 1.40 |
| 27 | Commercial | Under-report by 50% | From $t_{15}$ to $t_{35}$ | 1 | 0.50 | 1.10 | 0.48 |
| 32 | Commercial | Under-report by 60% | From $t_{15}$ to $t_{19}$ | 1.50 | 0.40 | 1.58 | 0.39 |
| | | Under-report by 50% | From $t_{35}$ to $t_{39}$ | 1 | 0.50 | 1.06 | 0.49 |
| 35 | Commercial | Under-report by 50% | All the time | 1 | 0.50 | 0.96 | 0.51 |
| 41 | Commercial | Over-report by 150% | All the time | −0.60 | 2.50 | −0.39 | 1.64 |
| Others | - | Honest | All the time | 0 | 1 | 0 | 1 |

that consumer 32 under-reports his energy consumption by 60% from 15-th to 19-th (since $\tilde{a}_{t_{15},32(EADF)}, \tilde{a}_{t_{16},32(EADF)}, \cdots, \tilde{a}_{t_{19},32(EADF)} > 0.05$, with $\overline{\tilde{a}}_{t_i,n(EADF)} = 1.58$) and reports his SM readings 50% less from 35-th to 39-th (since $\tilde{a}_{t_{35},32(EADF)}, \tilde{a}_{t_{36},32(EADF)} \cdots, \tilde{a}_{t_{39},32(EADF)} > 0.05$, with $\overline{\tilde{a}}_{t_i,n(EADF)} = 1.06$) time intervals. The results in Table 6.9 also show that Enhanced ADF is capable of revealing the amount of energy theft/loss (i.e., $\frac{1}{1+\overline{\tilde{a}}_{t_i,n(EADF)}}$) based on a small volume of consumers' power consumption data samples regardless of the presence of TLs/noise.

The results suggest that a detection rate of 100% was achieved by Enhanced ADF for the varying cheating/malfunctioning scenarios as setup in Tables 6.8 and 6.9 even in the presence of TLs and measurement noise.

## 6.4 Frameworks Validation Through AMI Test Rig

Since real-world SG energy theft samples and collector readings rarely, or do not, exist, regression and optimization analysis are first performed using the dataset released by Commission for Energy Regulation (2009) as discussed in Section 6.3. However, the dataset contains only consumers' energy consumption data. Therefore, the collector measurement $c_{t_i}$ is obtained by duly summing up the energy consumption of all consumers

University of Malaya

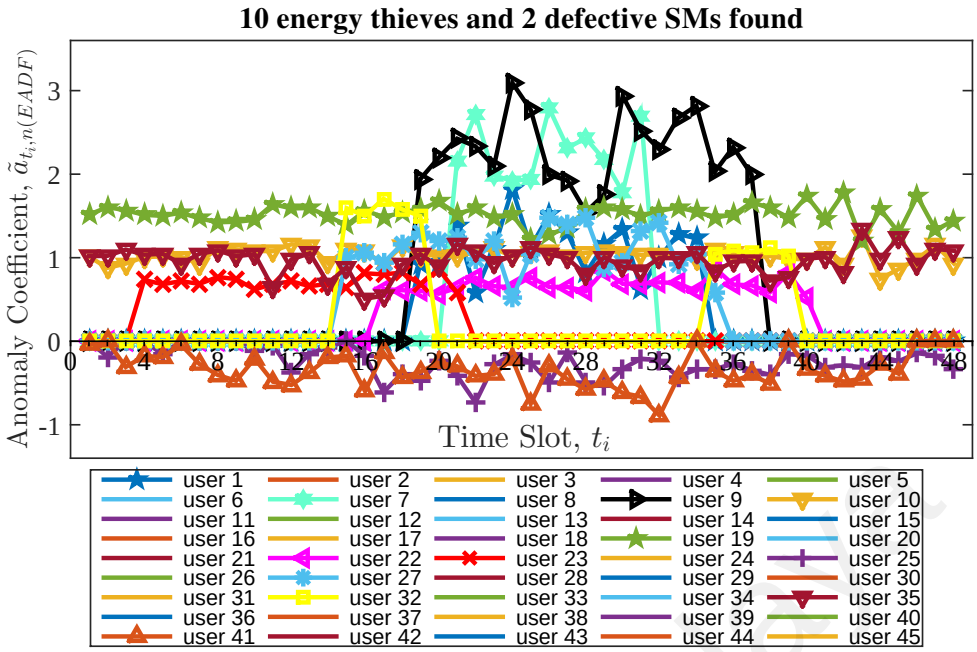



**Figure 6.9: Value of anomaly coefficients, $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced ADF when $a_{t_i,n}$ is varying (size of 45 consumers). Only anomalous cases are plotted.**

in the service area at time interval $t_i$ when there are no energy thefts and defective SMs. In this section, frameworks validation through AMI test rig, which consists of a DS and three consumers is conducted to validate the reliability and performance of the proposed LR-based and LP-based anomaly detection frameworks in real SG environment. The power line losses and the measurement noise/error of the equipment are taken into consideration in the hardware experimentation.

Similarly, two series of hardware experimentation which consist of both constant and varying scenarios are conducted on the 3-consumer test rig built in the laboratory. The hardware installation of the test rig is detailed in Section 5.2.2. Again, it is assumed that 30% of the consumers and/or SMs in the NAN have a non-zero anomaly coefficient. The minimum time of anomaly is subject to the time granularity of the SM (i.e., one slot = 30 minutes).

University of Malaya

### 6.4.1 Hardware Experimentation for LR-based Detection Framework

Recall that both the proposed LR-based detection schemes (i.e., LR-ETDM and CVLR-ETDM) do not consider the impact caused by TLs and measurement noise on the NTL detection analysis. In this section, the impact caused by real-world TLs and measurement noise on the performance of LR-based detection schemes are investigated and evaluated.

#### 6.4.1.1 Hardware Experimentation: LR-ETDM

The constant cheating/malfunctioning scenario for the test rig is setup as shown in Table 6.10. In particular, fraudulent consumers steal energy all the time and never stop cheating while some of the SMs had malfunctioned throughout the period of observations (i.e., case $s_2$ in the attack model). In Table 6.10, $a_n$ are exact settings from the test rig, whereas, $\tilde{a}_{n(LR)}$ are estimated values obtained by LR-ETDM.

Figure 6.10 depicts the estimated values of $\tilde{a}_{n(LR)}$ obtained from hardware experimentation. It can be observed from Table 6.10 that the $p$-values of all three anomaly coefficients (i.e., $\tilde{a}_{1(LR)}$, $\tilde{a}_{2(LR)}$ and $\tilde{a}_{3(LR)}$) are smaller than a 1‰ significance level. Thus, all three coefficients are shortlisted for further investigation. It is quite obvious from Figure 6.10 that consumer 2 is an energy thief (since $\tilde{a}_{2(LR)} > 0$) and the third SM is faulty (since $\tilde{a}_{3(LR)} < 0$). Then, based on these anomaly coefficients, the collector can compute the fraction of reported usage by $\frac{1}{1+\tilde{a}_{n(LR)}}$. Consumer 2 only reports 40% of his/her energy consumption since $\frac{1}{1+1.5046} \approx 0.40$, while consumer 3 over-reports what was consumed by 50% since $\frac{1}{1+(-0.3291)} \approx 1.50$. On the other hand, consumer 1 is classified as honest because $\frac{1}{1+0} \approx 1$.

The TLs in the test rig is approximately 0.15%, which might cause slight differences between the exact and estimated coefficients in Table 6.10. LR-ETDM is able to achieve a detection rate of 100% when the amounts of both TLs and measurement error in the test rig are insignificant.

University of Malaya

**Table 6.10: Comparison between $a_n$ and $\tilde{a}_{n(LR)}$ obtained by LR-ETDM from hardware experimentation**

| Consumer $n$ | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(LR)}$ | $p$-value$_{\tilde{a}}$ | $\frac{1}{1+\tilde{a}_{n(LR)}}$ |
|---|---|---|---|---|---|---|
| 1 | Honest | 0.00 | 1 | 0.0045≈ 0 | 0.0004 | 1 |
| 2 | Under-report by 60% | 1.50 | 0.40 | 1.5046 | 8.1259E-97 | 0.40 |
| 3 | Over-report by 50% | −0.3333 | 1.50 | −0.3291 | 2.2084E-87 | 1.50 |

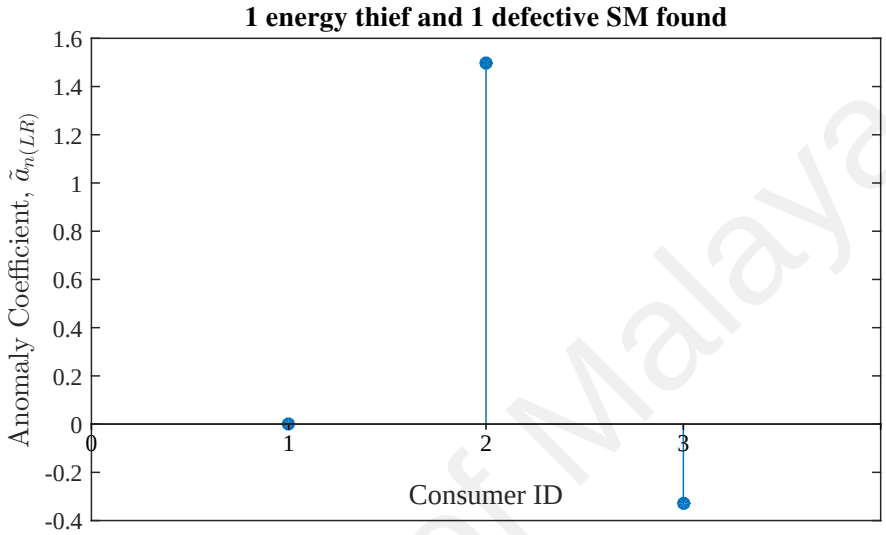



**Figure 6.10: Value of $\tilde{a}_{n(LR)}$ obtained by LR-ETDM from hardware experimentation (size of 3 consumers).**

### 6.4.1.2 Hardware Experimentation: CVLR-ETDM

Meanwhile, hardware experimentation is also conducted for the scenarios when energy thieves cheat on their energy reporting all the time/only during a certain period in a day (i.e., cases $s_2$, $s_3$ and $s_4$ in the attack model). The varying cheating/malfunctioning scenario for the test rig is setup as shown in Table 6.11. In the table, $a_n$ and $\beta_n$ are exact settings from the test rig, whereas, $\tilde{a}_{n(CVLR)}$ and $\tilde{\beta}_{n(CVLR)}$ are the estimated coefficients obtained by CVLR-ETDM. Recall that $a_n$ itself denotes the anomaly coefficients of consumers during off-peak period while $(a_n + \beta_n)$ denotes the anomaly coefficients of consumers during on-peak hours.

Figure 6.11 depicts the values of $\tilde{a}_{n(CVLR)}$ and $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ obtained from

University of Malaya

**Table 6.11: Comparison between $a_n$ & $\beta_n$ and $\tilde{a}_{n(CVLR)}$ & $\tilde{\beta}_{n(CVLR)}$ obtained by CVLR-ETDM from hardware experimentation**

| Consumer $n$ | Description | Affected Time Slot | $a_n$ | $\beta_n$ | $(a_n + \beta_n)$ | $\tilde{a}_{n(CVLR)}$ | $p$-value$_{\tilde{a}}$ | $\tilde{\beta}_{n(CVLR)}$ | $p$-value$_{\tilde{\beta}}$ | $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Under-report by 60% | All the time | 1.50 | 0.00 | 1.50 | 1.4870 | 1.9532E-72 | 0.0126≈ 0 | 0.0997 | 1.4996 |
| 2 | Under-report by 50% | On-peak | 0.00 | 1.00 | 1.00 | 0.0020 ≈ 0 | 0.2283 | 1.0076 | 2.1266E-71 | 1.0096 |
| 3 | Honest | All the time | 0.00 | 0.00 | 0.00 | 0.0044≈ 0 | 0.0141 | -0.0030≈ 0 | 0.3663 | 0.0015≈ 0 |

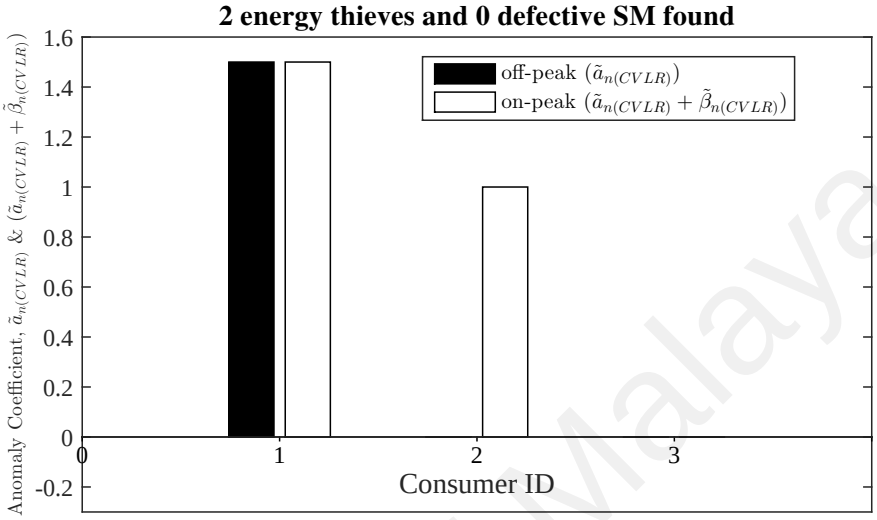



**Figure 6.11: Value of $\tilde{a}_{n(CVLR)}$ and $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ obtained by CVLR-ETDM from hardware experimentation (size of 3 consumers).**

hardware experimentation. Results in Figure 6.11 suggest that there are two dishonest consumers. In particular, since $\tilde{a}_{1(CVLR)} > 0$ and $\tilde{a}_{1(CVLR)} + \tilde{\beta}_{1(CVLR)} > 0$ (i.e., black and white bars co-appear), consumer 1 is classified as energy thief for stealing energy all the time. Meanwhile, consumer 2 reports less than what was consumed only during on-peak period because $\tilde{a}_{2(CVLR)} = 0$ and $\tilde{a}_{2(CVLR)} + \tilde{\beta}_{2(CVLR)} > 0$ (i.e., only white bar appears). Consumer 3 is always honest in energy reporting (since $\tilde{a}_{3(CVLR)} = 0$, $\tilde{a}_{3(CVLR)} + \tilde{\beta}_{3(CVLR)} = 0$). Similarly, the collector can compute the percentage of under-reporting/over-reporting based on the estimated $\tilde{a}_{n(CVLR)}$ and $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$, as discussed in Section 4.4.2. However, the computation for fraction of reported consumption of each consumer is omitted from Table 6.11 due to space constraints.

It can be seen from Table 6.11 that $p$-value of $\tilde{a}_{1(CVLR)}$ is less than a 1% significance

University of Malaya

level, implying consumer 1 is unlikely to be honest (i.e., $\tilde{a}_{1(CVLR)} \neq 0$). Meanwhile, $p$-values of both $\tilde{\beta}_{1(CVLR)}$ and $\tilde{\beta}_{3(CVLR)}$ are greater than 0.01, indicating that both consumer 1 and consumer 3 have $\tilde{\beta}_{n(CVLR)} = 0$ and hence *do not change* the cheating behaviors. Specifically, consumer 1 under-reports energy consumption all the time and consumer 3 is always honest. On the other hand, the $p$-value of $\tilde{\beta}_{2(CVLR)}$ is less than 0.01, indicating that consumer 2 is likely to have $\tilde{\beta}_{2(CVLR)} \neq 0$. A non-zero $\tilde{\beta}_{n(CVLR)}$ implies that the consumer has different cheating patterns throughout the period of observations. In such a case, consumer 2 only cheats during on-peak hours.

Similarly, the slight differences between the exact and estimated coefficients in Table 6.11 are possibly caused by the TLs/calibration errors in the test rig (i.e., only 0.15%). The experimentation result in Table 6.11 shows that CVLR-ETDM can attain a detection rate of 100% when the amounts of both TLs and measurement error are negligible.

### 6.4.2 Hardware Experimentation for LP-based Detection Framework

In pursuit of higher detection rate and lower false positives, metrics known as the *loss factor* and *error term* are introduced in the proposed LP-based anomaly detection framework to estimate the amount of TLs and capture the measurement noise, respectively in the distribution lines and transformers. The anomaly detection framework is then enhanced to detect consumers' malfeasance and faulty meters even when there are intermittent cheating and faulty equipment, improving its robustness. In this section, the performance and reliability of the LP-based framework are evaluated and validated in the presence of TLs and measurement noise of equipment in real SG environment.

#### 6.4.2.1 Hardware Experimentation: ADF

Here, the constant cheating/malfunctioning scenario for the test rig which is setup as shown in Table 6.12 is considered. One-day half-hourly energy data from the test rig are

University of Malaya

**Table 6.12: Comparison between constant $a_n$ and $\tilde{a}_{n(ADF)}$ obtained by ADF from hardware experimentation**

| Consumer $n$ | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(ADF)}$ | $\frac{1}{1+\tilde{a}_{n(ADF)}}$ |
|---|---|---|---|---|---|
| 1 | Honest | 0 | 1 | $-0.0079 \approx 0$ | 1 |
| 2 | Under-report by 60% | 1.5000 | 0.40 | 1.5000 | 0.40 |
| 3 | Over-report by 50% | −0.3333 | 1.50 | −0.3301 | 1.49 |

extracted for the theft detection analysis. The values of $a_n$ represent the exact state of each SM (i.e., honest, compromised or defective) from the test rig, whereas $\tilde{a}_{n(ADF)}$ denotes the computed anomaly coefficients obtained by ADF from the hardware experimentation.

Figure 6.12 depicts the estimated $\tilde{a}_{n(ADF)}$ under constant scenario. Results of $\tilde{a}_{n(ADF)}$ in Table 6.12 and Figure 6.12 suggest that the second consumer always under-reports what was consumed (since $\tilde{a}_{2(ADF)} > 0.05$) while the third SM over-reports the energy consumption (since $\tilde{a}_{3(ADF)} < -0.05$) all the time. Since $\tilde{a}_{1(ADF)} \approx 0$, consumer 1 is classified as honest. Based on the computed fraction of reported usage of each consumer in Table 6.12, it can be inferred that the second consumer only reports 40% of what was consumed and the third consumer over-reports what was consumed by 50%.

The TLs in the test rig is approximately 0%, which might cause inconsiderable differences between the exact and estimated anomaly coefficients. It is observed from Table 6.12 that the detection rate of ADF is 100% when the amounts of both TLs and measurement noise are trivial.

#### 6.4.2.2 Hardware Experimentation: Enhanced ADF

Subsequently, the varying cheating/malfunctioning scenario for the test rig which is setup as shown in Table 6.13 is considered. Four-day energy consumption data (i.e., $D = 4$) are extracted for solving the varying cheating problems. Here, the values of $a_{t_i,n}$ represent the exact state of each SM (i.e., honest, compromised or defective) from the test rig at each

University of Malaya

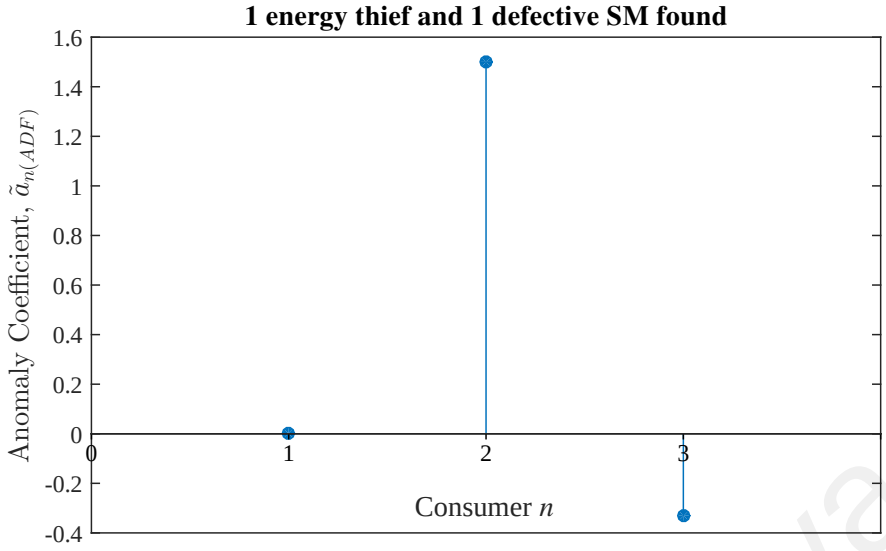



**Figure 6.12: Value of anomaly coefficients $\tilde{a}_{n(ADF)}$ obtained by ADF from hardware experimentation (size of 3 consumers).**

time interval $t_i$, whereas the values of $\overline{\tilde{a}}_{t_i,n(EADF)}$ are the average values of the estimated $\tilde{a}_{t_i,n(EADF)}$ during the affected time slots, whenever $\tilde{a}_{t_i,n(EADF)} \notin [-0.05, 0.05]$.

Figure 6.13 depicts the estimated $\tilde{a}_{t_i,n(EADF)}$ under varying scenario. As shown in Figure 6.13, it is observed that there are two consumers who have non-zero anomaly coefficients. In particular, since $\tilde{a}_{t_{17},1(EADF)}, \tilde{a}_{t_{18},1(EADF)}, \cdots, \tilde{a}_{t_{35},1(EADF)} > 0.05$, with $\overline{\tilde{a}}_{t_i,n(EADF)} = 0.9942$, the first consumer is classified as energy thief for stealing energy from the 17-th to the 35-th time slot. Meanwhile, the third energy thief under-reports what was consumed from 1-st to 15-th time slot (since $\tilde{a}_{t_1,3(EADF)}, \tilde{a}_{t_2,3(EADF)}, \cdots, \tilde{a}_{t_{15},3(EADF)} > 0.05$, with $\overline{\tilde{a}}_{t_i,n(EADF)} = 2.3558$) and from 27-th to 41-st time slot (since $\tilde{a}_{t_{27},3(EADF)}, \tilde{a}_{t_{28},3(EADF)}, \cdots, \tilde{a}_{t_{41},3(EADF)} > 0.05$, with $\overline{\tilde{a}}_{t_i,n(EADF)} = 2.3468$). The second consumer is honest in energy reporting because the estimated anomaly coefficient is zero. Similarly, the UPs can compute how much more/less the consumers have paid based on their computed fraction of reported consumption (i.e., $\frac{1}{1+\overline{\tilde{a}}_{t_i,n(EADF)}}$).

As mentioned earlier in Section 6.4.2.1, the TLs in the test rig is very small (i.e.,

University of Malaya

**Table 6.13: Comparison between varying $a_{t_i,n}$ and $\bar{\tilde{a}}_{t_i,n(EADF)}$ obtained by Enhanced ADF from hardware experimentation**

| Consumer $n$ | Description | Affected Time Slot, $t_i$ | $a_{t_i,n}$ | $\frac{1}{1+a_{t_i,n}}$ | $\bar{\tilde{a}}_{t_i,n(EADF)}$ | $\frac{1}{1+\bar{\tilde{a}}_{t_i,n(EADF)}}$ |
|---|---|---|---|---|---|---|
| 1 | Under-report by 50% | From $t_{17}$ to $t_{35}$ | 1 | 0.50 | 0.9942 | 0.50 |
| 2 | Honest | All the time | 0 | 1 | 0 | 1 |
| 3 | Under-report by 70% | From $t_1$ to $t_{15}$ | 2.33 | 0.30 | 2.3558 | 0.30 |
| | Under-report by 70% | From $t_{27}$ to $t_{41}$ | 2.33 | 0.30 | 2.3468 | 0.30 |

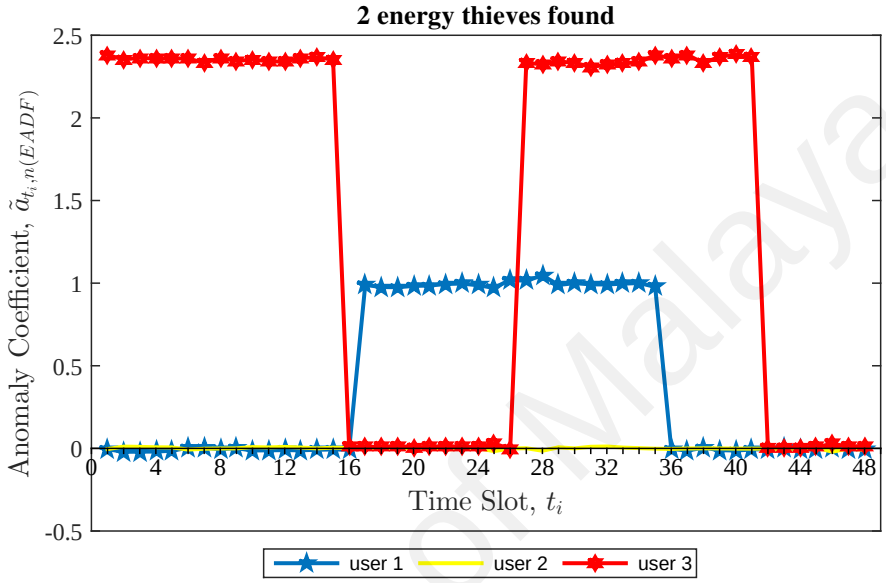



**Figure 6.13: Value of anomaly coefficients $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced ADF from hardware experimentation (size of 3 consumers).**

approximately 0%). The slight differences between the exact and estimated anomaly coefficients in Table 6.13 are likely due to TLs/measurement error. The results from Table 6.13 also suggest that the detection rate of Enhanced ADF is 100% when the amounts of both TLs and measurement error are nearly 0%.

## 6.5 Functional Comparison among NTL Detection Schemes

It is worth noting that both simulation-based and hardware experimentation-based comparative analyses between the proposed frameworks (i.e., LR-based and LP-based anomaly detection frameworks) and existing work are not made because they amounts to an unfair comparison. This is because some Artificial Intelligence (AI)-based energy theft detection schemes are susceptible to contamination attack and require large amount of

University of Malaya

training data (i.e., months and years) which might cause longer detection delay and limit the accuracy of the theft detection. Most of the other detection schemes will not perform satisfactorily with the smaller data sample size used for the proposed frameworks. In contrast, the proposed frameworks are not only robust against contamination attack but also able to reveal the amount of energy theft/loss based on a small volume of consumers' energy consumption data samples regardless of TLs/noise. Unlike some existing work such as the LUD-based energy theft detection scheme (S. Salinas et al., 2013), the proposed anomaly detection frameworks are not limited by the dimension of consumers' energy consumption data and they can still successfully identify all the fraudulent consumers and defective SMs in the NAN. Therefore, Table 6.14 *functionally* compares the proposed anomaly detection frameworks with existing NTL detection schemes, whereby the performance of the proposals is compared with the most recent and best results from existing energy theft detection schemes discussed in Section 2.4. It is observed that the detection rate of classification-based schemes (i.e., SVM (Nagi et al., 2010) and ELM techniques (Nizar et al., 2008)) are approximately $60\% - 70\%$. On the other hand, detection schemes adopting rough set theory (Spirić et al., 2014), Naïve Bayesian & Decision Tree (Nizar et al., 2007), LUD (S. Salinas et al., 2013), multi-class SVM (Jokar et al., 2016), LR (Yip et al., 2017)) and LP models (Yip, Tan, Tan, Gan, & Wong, 2018) yield higher detection rates, i.e., $\sim 100\%$. Although false positive rate captures how many honest consumers are wrongly classified as fraudulent ones by mistake (Jiang et al., 2014), only a few schemes such as (Nagi et al., 2010; Jokar et al., 2016) and (Krishna, Lee, et al., 2016) reported the quantitative false positive rate. In addition, most NTL detection schemes do not consider TLs, which may prohibit their deployment for actual utilization. In contrast, to make the proposed framework more practical, the loss factor and error term are introduced in the LP-based detection framework to estimate the percentage of TLs and capture the

University of Malaya

**Table 6.14: Comparison among energy theft detection schemes**

| Category | Scheme | Technique in Use | Detection Rate (%) | False Positive (%) | Consider TLs | Detect Over-reporting by SM |
|---|---|---|---|---|---|---|
| Conventional Power Grids | Ref. (Nagi et al., 2010) | SVM | 60 | 13.57 | ✗ | ✗ |
| | Ref. (Nizar et al., 2008) | ELM & Online Sequential-ELM | 70 | - | ✗ | ✗ |
| | Ref. (Nizar et al., 2007) | Naïve Bayesian & Decision Tree | 99 | - | ✗ | ✗ |
| | Ref. (Dos Angelos et al., 2011) | Fuzzy Clustering & Classification | 80 | - | ✗ | ✗ |
| | Ref. (Spirić et al., 2014) | Rough Set Theory | 93 | - | ✗ | ✗ |
| Smart Grids | Ref. (S. Salinas et al., 2013) | LUD | 100 | - | ✗ | ✓ |
| | Ref. (Jokar et al., 2016) | Multi-class SVM | 94 | 11 | ✗ | ✗ |
| | Ref. (Krishna, Lee, et al., 2016) | Kullback-Leibler Divergence | 82 | 0 | ✗ | ✗ |
| | LR-ETDM & CVLR-ETDM | Linear Regression | 100 | 0 | ✗ | ✓ |
| | ADF & Enhanced ADF | Linear Programming | 100 | 0 | ✓ | ✓ |

measurement noise, respectively. Despite the introduction of these terms, the proposed framework is still able to correctly detect pilfering of electricity and defective SMs. In addition to detecting SMs that under-report (i.e., energy theft), LUD model (S. Salinas et al., 2013) and the proposed models (i.e., LR-ETDM & CVLR-ETDM and ADF & Enhanced ADF) are able to identify SMs which over-report the energy consumption. Furthermore, the LP-based detection framework outperforms the LR-based detectors in handling NTLs because the energy thefts/meter irregularities can be detected regardless of whether they occur all the time or at varying rates during intermittent periods in a day. Consumers' anomaly coefficient at each time slot is evaluated separately and hence Enhanced ADF is still able to accurately detect pilfering of electricity and defective SMs even though NTLs take place at irregular intervals. Therefore, the proposed LP-based anomaly detection framework is more robust.

## 6.6 Impact of Distributed Energy Resources on the Frameworks

In recent years, many countries have been actively encouraging the adoption of renewable energy/DER to replace fossil fuels. For example, in Malaysia, the DER (i.e., solar panels, wind turbines and etc.) feedings method can be grouped under two categories, namely direct feed and indirect feed, respectively (Abu, H. A., Saharuddin, S., Hussein, Z. F., Malathy, B., Busrah, A. M., & Devaraju, P., 2013). The connection of direct feed and indirect feed are shown in Figures 6.14 and 6.15, respectively.

University of Malaya

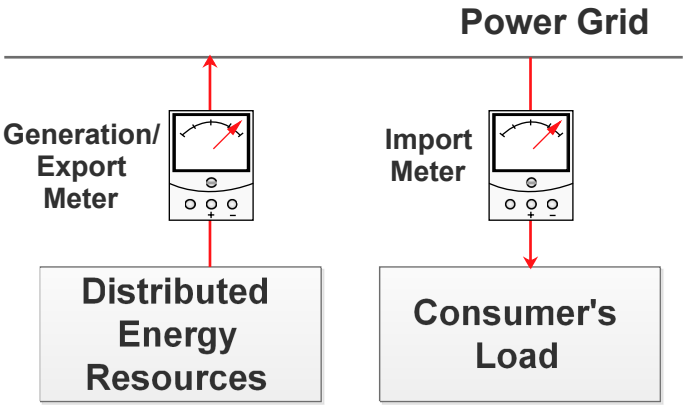



**Figure 6.14: Connection to power grid (direct).**

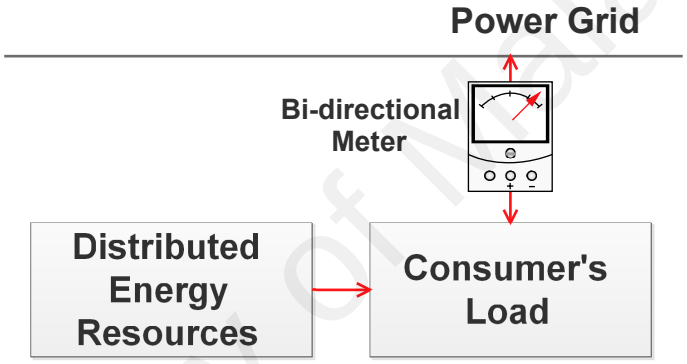



**Figure 6.15: Connection to power grid (indirect).**

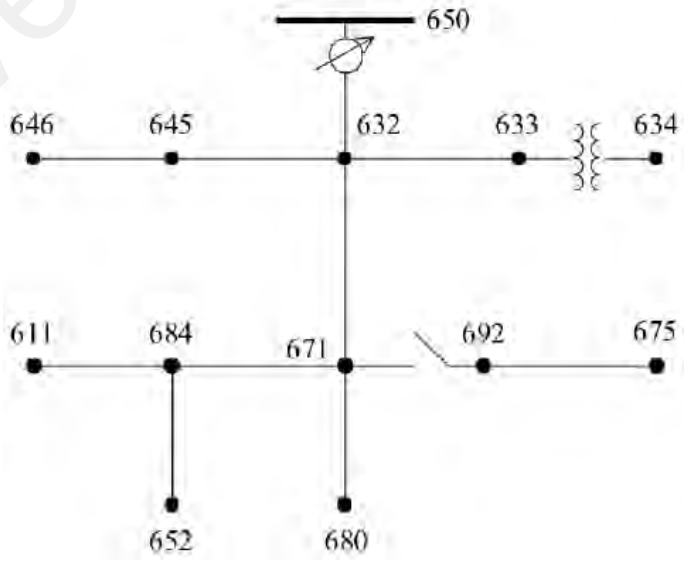



**Figure 6.16: Modified IEEE 13-node test feeder**

University of Malaya

For direct feeding, the generated energy is fed directly to power grid, as shown in Figure 6.14. A generation meter/export meter will record the amount of energy generated and fed into the power grid. UPs will pay the owner of DER based on the reading from generation meter. Meanwhile, for indirect feeding, the SM will perform a net metering function which allows renewable power generated to be used at the home/building before any excess energy is fed back into the grid, as illustrated in Figure 6.15. The bi-directional SM can measure how much energy is flowing in the opposite direction, back into the grid. Indirect connection is allowed for special case and requires additional verification and supplementary agreement with the UPs.

UPs incorporating DER have to plan for new connections and business model to achieve accurate control and forecasting needed for grid reliability and security. Therefore, a slight modification can be done on Equation (3.1) to take into account the impact of DER on the frameworks. In this work, the focus is on the detection of under-reporting and over-reporting of consumers' SM (i.e., import meter), thereby it is assumed that the DER generation measurements recorded by the generation meter are not manipulated. The detection of malicious consumers who over-report the energy they generate for financial gain may be considered in a future work. According to the modified IEEE 13-node test feeder as shown in Figure 6.16 (IEEE Power & Energy Society, 1992), the total energy supplied by the UP to the NAN ($c_{t_i}$) and the total energy generated by all the consumers at time interval $t_i$ ($g_{t_i}$) (i.e., root node) should tally with the sum of electricity consumption reported by all the consumers (i.e., leaf nodes) as discussed in Section 3.3.2. Therefore, the following equation is formulated:

$$c_{t_i} + g_{t_i} = \sum_{n=1}^{N} p_{t_i,n} + \lambda + \theta + \gamma, \tag{6.4}$$

where $\lambda$ denotes the TLs, while $\theta$ and $\gamma$ represent the inaccurate meter readings due to



University of Malaya

**Table 6.15: Comparison among constant $a_n$, $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ obtained from hardware experimentation**

| Consumer $n$ | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(LR)}$ | $\frac{1}{1+\tilde{a}_{n(LR)}}$ | $\tilde{a}_{n(ADF)}$ | $\frac{1}{1+\tilde{a}_{n(ADF)}}$ |
|---|---|---|---|---|---|---|---|
| 1 | Over-report by 50% | -0.3333 | 1.50 | -0.3366 | 1.51 | -0.3386 | 1.51 |
| 2 | Honest | 0 | 1 | $0.0018 \approx 0$ | 1 | 0 | 1 |
| 3 | Under-report by 60% | 1.50 | 0.40 | 1.5167 | 0.40 | 1.5119 | 0.40 |

energy frauds and defective SMs, respectively.

## 6.7 Strengths and Weaknesses of the Proposed Frameworks

LR-based and LP-based anomaly detection frameworks are put forward to study consumers' energy consumption behavior for detecting the localities of metering defects as well as energy thefts. In this section, comparison studies are performed on the proposed anomaly detection frameworks in real SG environment. To investigate the scalability of both proposed frameworks, data are also extracted from the Irish Smart Energy Trial in the performance comparison studies.

### 6.7.1 Constant Anomaly Coefficients

Here, the performance of LR-ETDM and ADF where the dishonest consumers always under-report their energy consumption (i.e., case $s_2$ of attack model where $\nu \in (0, 0.95)$) and/or defective SMs over-report what was consumed (i.e., case $s_2$ where $\nu \in (1.05, 2.5]$) throughout the entire day are compared. The constant under-reporting/over-reporting scenario for the size of three consumers of the test rig is setup as shown in Table 6.15. Daily half-hourly smart energy data (i.e., 48 data points) from the AMI test rig are extracted for the anomaly detection analysis. Values of $a_n$ depict the exact state of each SM (i.e., honest, compromised or faulty) from the dataset, whereas, $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ are the anomaly coefficients obtained by the LR-ETDM and ADF, respectively.

Figure 6.17 depicts the values of $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ when $a_n$ is constant for the size of

University of Malaya

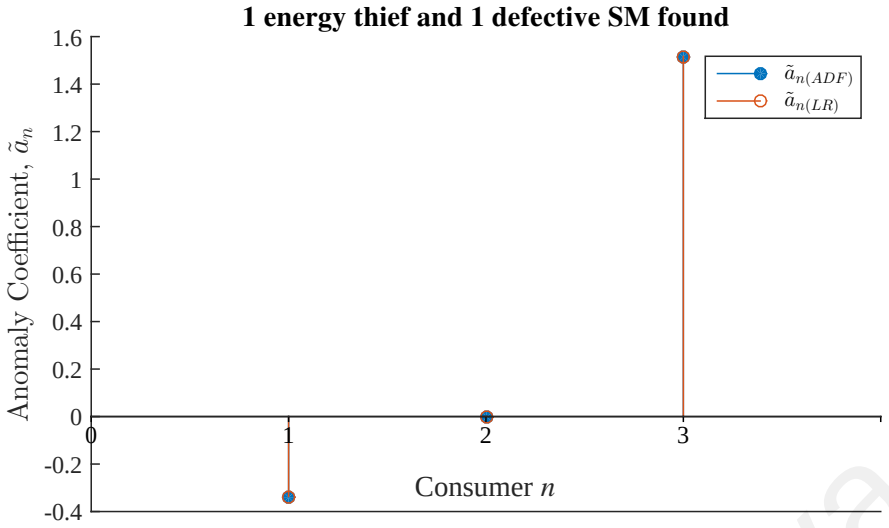



**Figure 6.17: Values of anomaly coefficients obtained by LR-ETDM and ADF from the test rig when $a_n$ is constant (size of 3 consumers).**

three consumers. It can be observed from Table 6.15 and Figure 6.17 that the first SM is out of order as $\tilde{a}_1 < -0.05$ while the third consumer is an energy thief since $\tilde{a}_3 > 0.05$. Consumer 2 is classified as honest as $\tilde{a}_2 = 0$. The results also suggest that both $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ achieve detection rate of 100% and show similar results when the service area is small and the amounts of both TLs and measurement error are negligible.

To study how the proposed frameworks scale with the number of consumers in the event of TLs and measurement noise being present, a NAN of 45 consumers is considered. The simulation for the case of 45 consumers is setup in a similar mean as shown in Table 6.16. Four-day half-hourly smart energy data (i.e., 192 data points) from the Irish Smart Energy Trial (Commission for Energy Regulation, 2009) are extracted for the detection analysis.

As mentioned earlier, the average TLs of LV network in Malaysia was reported to range from 0.59% to 3.23%. To show the viability of the proposed anomaly detection frameworks in estimating the amount of TLs, an evaluation environment with $3\% - 5\%$ of TLs is created, as presented in Equation (6.2). Recall that, the state-of-the-art SMs measure very accurate measurements, where the errors are usually modeled by white uncorrelated

University of Malaya

**Table 6.16: Comparison among constant $a_n$, $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ obtained from the Irish Smart Energy Trial (size of 45 consumers)**

| Consumer $n$ | Description | $a_n$ | $\frac{1}{1+a_n}$ | $\tilde{a}_{n(LR)}$ | $\frac{1}{1+\tilde{a}_{n(LR)}}$ | $\tilde{a}_{n(ADF)}$ | $\frac{1}{1+\tilde{a}_{n(ADF)}}$ |
|---|---|---|---|---|---|---|---|
| 1 | Over-report by 70% | −0.4117 | 1.7 | −0.3275 | 1.49 | −0.4181 | 1.72 |
| 7 | Over-report by 50% | −0.3333 | 1.50 | −0.2750 | 1.38 | −0.3144 | 1.46 |
| 10 | Under-report by 60% | 1.5000 | 0.40 | 1.5533 | 0.39 | 1.5013 | 0.40 |
| 17 | Under-report by 25% | 0.3333 | 0.75 | 0.4142 | 0.71 | 0.3193 | 0.76 |
| 19 | Honest | 0 | 1 | *0.0920 | 0.92 | 0 | 1 |
| 22 | Under-report by 70% | 2.3333 | 0.30 | 2.5132 | 0.28 | 2.3381 | 0.30 |
| 24 | Honest | 0 | 1 | *0.1331 | 0.88 | 0 | 1 |
| 27 | Under-report by 70% | 2.3333 | 0.30 | 2.4003 | 0.29 | 2.3064 | 0.30 |
| 31 | Over-report by 100% | −0.5000 | 2.00 | −0.4647 | 1.87 | −0.4985 | 1.99 |
| 32 | Honest | 0 | 1 | $0.0501 \approx 0$ | 1 | 0 | 1 |
| 34 | Honest | 0 | 1 | *0.0988 | 0.91 | 0 | 1 |
| 35 | Under-report by 65% | 1.8571 | 0.35 | 2.0233 | 0.33 | 1.8549 | 0.35 |
| 38 | Under-report by 50% | 1 | 0.50 | 1.1801 | 0.46 | 1.0344 | 0.49 |
| 41 | Over-report by 50% | −0.3333 | 1.50 | *0 | 1 | −0.2545 | 1.34 |
| 42 | Under-report by 30% | 0.4285 | 0.70 | 0.3491 | 0.74 | 0.3885 | 0.72 |
| 45 | Under-report by 30% | 0.4285 | 0.70 | 0.5508 | 0.64 | 0.4454 | 0.69 |
| Others | Honest | 0 | 1 | - | - | - | - |

* False Positive

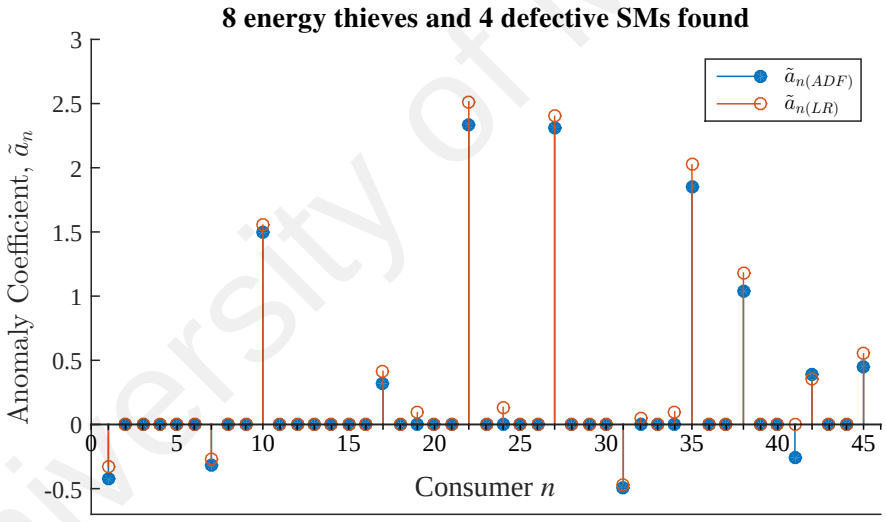



**Figure 6.18: Values of anomaly coefficients obtained by LR-ETDM and ADF from the Irish Smart Energy Trial when $a_n$ is constant (size of 45 consumers).**

noise with standard deviation and zero mean (S. A. Salinas & Li, 2015). Thus, to further validate the proposed frameworks, the measurement errors are also considered as shown in Equation (6.3).

As shown in Table 6.16 and Figure 6.18, LR-ETDM tends to produce inaccurate anomaly coefficient vector in the presence of TLs and measurement noise. Specifically, some of

University of Malaya

the honest consumers are accused wrongly as fraudulent consumers (i.e., consumers 19, 24 and 34) while consumer 41 who over-reports his/her energy consumption is classified incorrectly as normal. As discussed earlier, any SMs who have anomaly coefficients in $[-0.05, 0.05]$ are assumed to be truthful in energy reporting (i.e., $\tilde{a}_n \approx 0$). Therefore, consumer 32 is identified as honest. In contrast, ADF produces more accurate anomaly coefficients and is able to detect all anomalous consumers and faulty SMs successfully after taking into consideration the impact caused by TLs and measurement noise.

Then, based on the computed anomaly coefficients, the operation center can compute the fraction of reported energy usage of each consumer by $\frac{1}{1+\tilde{a}}$ as shown in Tables 6.15 and 6.16.

### 6.7.2 Varying Anomaly Coefficients

Besides, performance comparison between CVLR-ETDM and Enhanced ADF for the case where there are anomalous fraudsters who cheat on their energy reporting (i.e., case $s_3$ of attack model where $\nu \in (0, 0.95)$ or case $s_4$) and/or defective SMs over-report what was consumed (i.e., case $s_3$ where $\nu \in (1.05, 2.5]$) only during off-peak/on-peak periods is also conducted. The varying under-reporting/over-reporting scenario for the hardware experimentation (size of 3 consumers) is setup as shown in Table 6.17.

Here, values of $a_{t_i,n}$ are exact settings from the dataset, whereas, $\tilde{a}_{n(CVLR)}$ and $\tilde{\beta}_{n(CVLR)}$ are the estimated values obtained by the CVLR-ETDM. Values of $\overline{\tilde{a}}_{t_i,n(EADF)}$ are the average values of the computed $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced ADF during the affected time slots, whenever $\tilde{a}_{t_i,n(EADF)} \notin [-0.05, 0.05]$. One-day and four-day energy consumption data are extracted for solving the varying cheating problems using CVLR-ETDM and Enhanced ADF, respectively. Four-day metered data are needed in Enhanced ADF as it requires at least $N$ days data for detection analysis (i.e., $N$ is the number of consumers in the service area). Figures 6.19(a) and 6.19(b) depict the values of $\tilde{a}_{n(CVLR)}$ and $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$

University of Malaya

obtained by CVLR-ETDM as well as the values of $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced ADF, respectively, for the hardware experimentation which is setup as shown in Table 6.17.

As shown in Figures 6.19(a) and 6.19(b), it is obvious that consumer 1 over-reports his/her energy consumption all the time. Particularly, $\tilde{a}_{1(CVLR)} < -0.05$, $(\tilde{a}_{1(CVLR)} + \tilde{\beta}_{1(CVLR)}) < -0.05$ while $(\tilde{a}_{t_1,1(EADF)}, \tilde{a}_{t_2,1(EADF)}, \cdots, \tilde{a}_{t_{48},1(EADF)} < -0.05$, where $\overline{\tilde{a}}_{t_i,1(EADF)} = -0.3371)$. On the other hand, consumer 2 under-reports what was consumed only during on-peak period (i.e., $\tilde{a}_{2(CVLR)} = 0$, $(\tilde{a}_{2(CVLR)} + \tilde{\beta}_{2(CVLR)}) > 0.05$ while $(\tilde{a}_{t_{16},2(EADF)}, \tilde{a}_{t_{17},2(EADF)}, \cdots, \tilde{a}_{t_{39},2(EADF)} > 0.05$, where $\overline{\tilde{a}}_{t_i,2(EADF)} = 1.5001))$. Consumer 3 is identified as honest as $\tilde{a}_{3(CVLR)} = 0$, $(\tilde{a}_{3(CVLR)} + \tilde{\beta}_{3(CVLR)}) = 0$ while $(\tilde{a}_{t_1,3(EADF)}, \tilde{a}_{t_2,3(EADF)}, \cdots, \tilde{a}_{t_{48},3(EADF)} = 0)$. Similarly, both CVLR-ETDM and Enhanced ADF produce almost the same results when the amounts of both TLs and measurement error are negligible.

To assess the scalability of the proposed anomaly detection frameworks under varying cheating/malfunctioning scenario in the presence of TLs and measurement noise, a NAN of 45 consumers is considered. Five-month half-hourly smart energy data (i.e., $D = 150$) from the Irish Smart Energy Trial (Commission for Energy Regulation, 2009) are extracted for the detection analysis. The varying cheating/malfunctioning simulation for the scenario of 45 consumers in the presence of TLs and noise is setup as shown in Table 6.18.

By referring to the results from Figures 6.20(a) and 6.20(b), it can be inferred that there are eleven fraudulent consumers and a defective SM. Specifically, consumers 1, 19 and 42 steal energy all the time (i.e., $\tilde{a}_{n(CVLR)} > 0.05$, $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)}) > 0.05$ while $(\tilde{a}_{t_1,n(EADF)}, \tilde{a}_{t_2,n(EADF)}, \cdots, \tilde{a}_{t_{48},n(EADF)} > 0.05)$. Consumer 37 always over-reports what was consumed (i.e., $\tilde{a}_{37(CVLR)} < -0.05$, $(\tilde{a}_{37(CVLR)} + \tilde{\beta}_{37(CVLR)}) < -0.05$ while $(\tilde{a}_{t_1,37(EADF)}, \tilde{a}_{t_2,37(EADF)}, \cdots, \tilde{a}_{t_{48},37(EADF)} < -0.05)$. Meanwhile, consumer 8 under-reports what was consumed only during off-peak period (i.e., $\tilde{a}_{8(CVLR)} > 0.05$,

University of Malaya

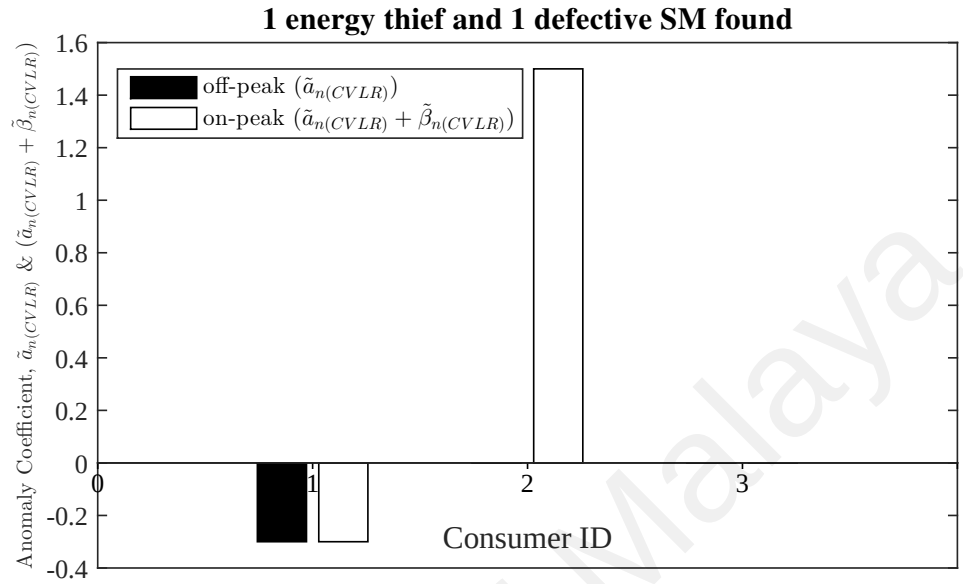
$(\tilde{a}_{8(CVLR)} + \tilde{\beta}_{8(CVLR)}) = 0$, while $(\tilde{a}_{t_1,8(EADF)}, \tilde{a}_{t_2,8(EADF)}, \cdots, \tilde{a}_{t_{15},8(EADF)} > 0.05$ and $\tilde{a}_{t_{40},8(EADF)}, \tilde{a}_{t_{41},8(EADF)}, \cdots, \tilde{a}_{t_{48},8(EADF)} > 0.05)$. As shown in Figure 6.20(a), it is observed that all the honest consumers do not have $\tilde{a}_n = 0$ or $(\tilde{a}_n + \tilde{\beta}_n) = 0$. The slight errors are due to the injected TLs and noise in Equation (6.3). Meanwhile, as presented in Table 6.18, the combination of $\tilde{a}_{4(CVLR)} \approx 0$ and $(\tilde{a}_{4(CVLR)} + \tilde{\beta}_{4(CVLR)}) = 0.10$ indicates that consumer 4 steals energy only during on-peak period. In addition, the combination of $\tilde{a}_{38(CVLR)} > 0$ and $(\tilde{a}_{38(CVLR)} + \tilde{\beta}_{38(CVLR)}) > 0$ indicates that consumer 38 steals energy all the time. However, in actual experimentation, consumer 4 is honest and consumer 38 under-reports what was consumed only during off-peak period. These results suggest that CVLR-ETDM becomes unstable in the presence of TLs and calibration noise in larger service area. On the contrary, Enhanced ADF is capable of identifying the anomalous and faulty SMs accurately under this scenario. As mentioned previously, the operation center can calculate the fraction of reported energy usage of each consumer by $\frac{1}{1+a}$ or $\frac{1}{1+(a+\beta)}$ (i.e., whenever $\tilde{a}_n \approx 0$) as shown in Tables 6.17 and 6.18 based on the computed coefficients.

University of Malaya

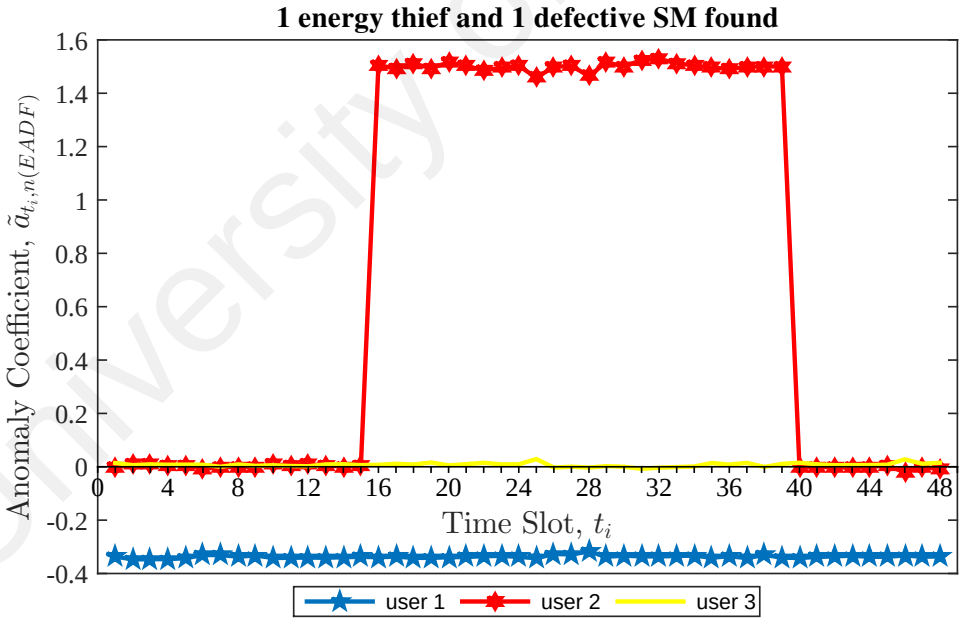
**Table 6.17: Comparison among varying $a_{t_i,n}$, $\tilde{a}_{n(CVLR)}$, $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ and $\overline{\tilde{a}}_{t_i,n(EADF)}$ obtained from hardware experimentation**

| Consumer $n$ | Description | Affected Time Slot, $t_i$ | $a_{t_i,n}$ | $\frac{1}{1+a_{t_i,n}}$ | $\tilde{a}_{n(CVLR)}$ | $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ | $\frac{1}{1+\tilde{a}_{n(CVLR)}}$ or $\frac{1}{1+(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})}$ | $\overline{\tilde{a}}_{t_i,n(EADF)}$ | $\frac{1}{1+\overline{\tilde{a}}_{t_i,n(EADF)}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Over-report by 50% | All the time | −0.3333 | 1.50 | −0.3368 | −0.3335 | 1.51 | −0.3371 | 1.51 |
| 2 | Under-report by 60% | On-peak (From $t_{16}$ to $t_{39}$) | 1.5000 | 0.40 | 0.0019851 ≈ 0 | 1.5120 | 0.40 | 1.5001 | 0.40 |
| 3 | Honest | All the time | 0 | 1 | 0.0044069 ≈ 0 | 0.0014501≈ 0 | 1 | 0 | 1 |

University of Malaya



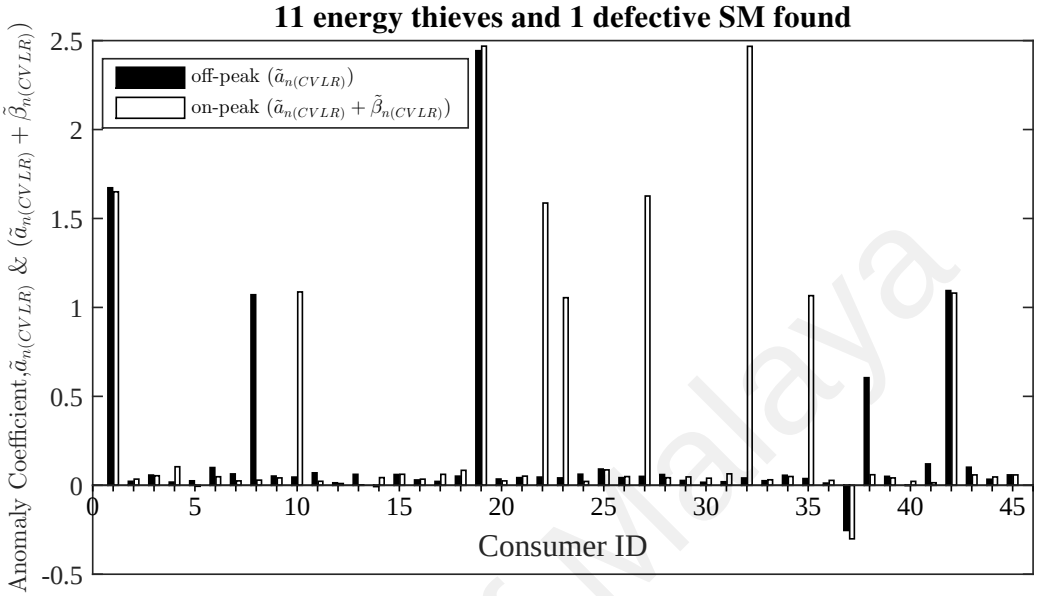

(a) CVLR-ETDM



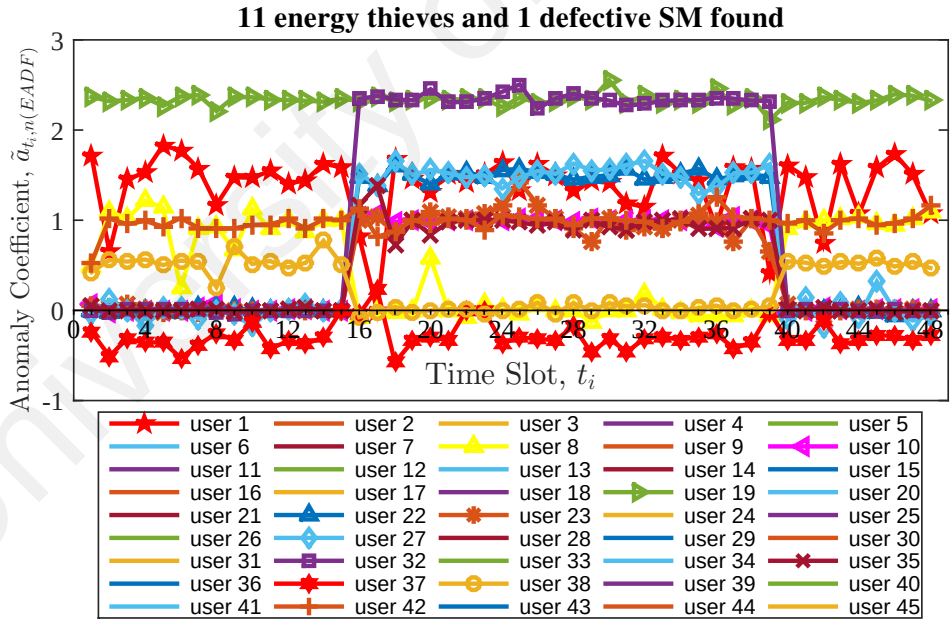

(b) Enhanced ADF

**Figure 6.19: Values of anomaly coefficients obtained by CVLR-ETDM and Enhanced ADF from the test rig when $a_{t_i,n}$ is varying (size of 3 consumers)**

University of Malaya




(a) CVLR-ETDM




(b) Enhanced ADF

**Figure 6.20: Values of anomaly coefficients obtained by CVLR-ETDM and Enhanced ADF from the Irish Smart Energy Trial when $a_{t_i,n}$ is varying (size of 45 consumers)**

University of Malaya

**Table 6.18: Comparison among varying $a_{t_i,n}$, $\tilde{a}_{n(CVLR)}$, $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ and $\overline{\tilde{a}}_{t_i,n(EADF)}$ obtained from the Irish Smart Energy Trial (size of 45 consumers)**

| Consumer $n$ | Description | Affected Time Slot, $t_i$ | $a_{t_i,n}$ | $\frac{1}{1+a_{t_i,n}}$ | $\tilde{a}_{n(CVLR)}$ | $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ | $\frac{1}{1+\tilde{a}_{n(CVLR)}}$ or $\frac{1}{1+(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})}$ | $\overline{\tilde{a}}_{n(EADF)}$ | $\frac{1}{1+\overline{\tilde{a}}_{n(EADF)}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Under-report by 60% | All the time | 1.5000 | 0.40 | 1.6726 | 1.6501 | 0.38 | 1.4870 | 0.40 |
| 4 | Honest | All the time | 0 | 1 | $0.0179 \approx 0$ | *0.1042 | 0.91 | 0 | 1 |
| 8 | Under-report by 50% | Off-peak (From $t_1$ to $t_{15}$), (From $t_{40}$ to $t_{48}$) | 1 | 0.50 | 1.0713 | $0.029 \approx 0$ | 0.48 | 0.9509 | 0.51 |
| 10 | Under-report by 50% | On-peak (From $t_{16}$ to $t_{39}$) | 1 | 0.50 | $0.0461 \approx 0$ | 1.0871 | 0.48 | 1.0000 | 0.50 |
| 19 | Under-report by 70% | All the time | 2.3333 | 0.30 | 2.4433 | 2.4689 | 0.29 | 2.3306 | 0.30 |
| 22 | Under-report by 60% | On-peak (From $t_{16}$ to $t_{39}$) | 1.5000 | 0.40 | $0.0461 \approx 0$ | 1.5869 | 0.39 | 1.4995 | 0.40 |
| 23 | Under-report by 50% | On-peak (From $t_{16}$ to $t_{39}$) | 1 | 0.50 | $0.0415 \approx 0$ | 1.0547 | 0.49 | 0.9997 | 0.50 |
| 27 | Under-report by 60% | On-peak (From $t_{16}$ to $t_{39}$) | 1.5000 | 0.40 | $0.0500 \approx 0$ | 1.6263 | 0.38 | 1.5209 | 0.40 |
| 32 | Under-report by 70% | On-peak (From $t_{16}$ to $t_{39}$) | 2.3333 | 0.30 | $0.0411 \approx 0$ | 2.4679 | 0.29 | 2.3250 | 0.30 |
| 35 | Under-report by 50% | On-peak (From $t_{16}$ to $t_{39}$) | 1 | 0.50 | $0.0378 \approx 0$ | 1.0662 | 0.48 | 0.9977 | 0.50 |
| 37 | Over-report by 50% | All the time | −0.3333 | 1.50 | −0.2539 | −0.3010 | 1.43 | −0.3354 | 1.50 |
| 38 | Under-report by 35% | Off-peak (From $t_1$ to $t_{15}$), (From $t_{40}$ to $t_{48}$) | 0.5385 | 0.65 | 0.6051 | *0.0595 | 0.62 | 0.5433 | 0.65 |
| 42 | Under-report by 50% | All the time | 1 | 0.50 | 1.0946 | 1.0806 | 0.48 | 0.9974 | 0.50 |
| Others | Honest | All the time | 0 | 1 | - | - | - | - | |

* False Positive

University of Malaya

### 6.7.3 Performance Comparison Between LR-based and LP-based Anomaly Detection Frameworks

Table 6.19 shows the performance comparison between LR-ETDM and ADF in detecting the constant cheating/malfunctioning under different scenarios on the same dataset. It can be observed from Table 6.19 that the detection rates (DR) of both LR-ETDM and ADF are 100% when the percentage of TLs is negligible (i.e., case 1) or when TLs are non-existent (i.e., case 2). The results also suggest that the detection of both schemes becomes more accurate when the consumers' energy consumption data are observed over longer periods. This is due to the fact that observation of metered data over an extended period of time results in addition in the number of constraints which can enhance the accuracy of the theft detection analysis. false positives (FP) is an important metric which indicates how many honest consumers are classified into malicious ones by mistake. Although LR-ETDM achieves higher detection accuracy when the metered data are observed over longer times, the number of FP increases because the impact caused by TLs and noise on the detection analysis is not considered in the framework. On the contrary, the number of FP decreases in ADF when consumers' energy consumption data are observed over more days. In ADF, loss factor and error term capture the percentage of TLs and calibration errors at each time interval, respectively. Therefore, it provides a more robust detection as compared to LR-ETDM.

On the other hand, Table 6.20 demonstrates the comparison studies between CVLR-ETDM and Enhanced ADF in detecting the varying cheating/equipment malfunctioning under different scenarios on the same data sample. Similarly, detection rates of both CVLR-ETDM and Enhanced ADF are 100% when the amount of TLs is small (i.e., cases 9 and 10) or when TLs are non-existent (i.e., cases 11 and 18). It can be seen from Table 6.20 that Enhanced ADF requires more observation data (i.e., at least $N$ days, where $N$ is the

University of Malaya

**Table 6.19: Performance comparison between LR-ETDM and ADF**

| Scenario | No. of Consumers | No. of Days | In the Presence of TLs | LR-ETDM | | ADF | |
|---|---|---|---|---|---|---|---|
| | | | | DR (%) | No. of FP | DR (%) | No. of FP |
| 1 | 3 (test rig) | 1 | < 1% | 100 | 0 | 100 | 0 |
| 2 | 45 | 1 | ✗ | 100 | 0 | 100 | 0 |
| 3 | 45 | 1 | ✓ | 16.67 | 0 | 66.67 | 18 |
| 4 | 45 | 2 | ✓ | 83.33 | 2 | 91.67 | 9 |
| 5 | 45 | 3 | ✓ | 91.67 | 2 | 91.67 | 7 |
| 6 | 45 | 4 | ✓ | 91.67 | 3 | 100 | 0 |
| 7 | 45 | 5 | ✓ | 91.67 | 3 | 100 | 0 |
| 8 | 45 | 6 | ✓ | 100 | 4 | 100 | 0 |

DR: detection rate; FP: false positives

**Table 6.20: Performance comparison between CVLR-ETDM and Enhanced ADF**

| Scenario | No. of Consumers | No. of Days | In the Presence of TLs | CVLR-ETDM | | Enhanced ADF | |
|---|---|---|---|---|---|---|---|
| | | | | DR (%) | No. of FP | DR (%) | No. of FP |
| 9 | 3 (test rig) | 1 | < 1% | 100 | 0 | - | - |
| 10 | 3 (test rig) | 4 | < 1% | 100 | 0 | 100 | 0 |
| 11 | 45 | 2 | ✗ | 100 | 0 | - | - |
| 12 | 45 | 2 | ✓ | 8.33 | 0 | - | - |
| 13 | 45 | 3 | ✓ | 75 | 0 | - | - |
| 14 | 45 | 4 | ✓ | 75 | 0 | - | - |
| 15 | 45 | 5 | ✓ | 75 | 2 | - | - |
| 16 | 45 | 6 | ✓ | 83.33 | 3 | - | - |
| 17 | 45 | 30 | ✓ | 100 | 0 | - | - |
| 18 | 45 | 45 | ✗ | 100 | 0 | 100 | 0 |
| 19 | 45 | 45 | ✓ | 100 | 0 | 91.67 | 1 |
| 20 | 45 | 60 | ✓ | 100 | 1 | 100 | 0 |
| 21 | 45 | 90 | ✓ | 100 | 1 | 100 | 0 |
| 22 | 45 | 150 | ✓ | 100 | 1 | 100 | 0 |

DR: detection rate; FP: false positives

number of consumers in the service area) as compared to CVLR-ETDM to achieve higher detection rate (DR) and lower FP. The table also suggests that both frameworks obtain higher DR when the metered data are observed over longer periods. However, the number of FP increases over time in CVLR-ETDM as the effect of TLs is not considered. On the other hand, the number of FP reduces over time in Enhanced ADF as loss factor and error term capture the percentage of TLs and noise in the system, respectively, thereby improving the DR.

Although LR-ETDM and CVLR-ETDM are able to detect the energy thieves and faulty SMs more accurately when the metered data are observed over an extended period of

University of Malaya

**Table 6.21: Summary of the proposed frameworks**

| Scheme | Technique In Use | Detect Constant Cheating/ Malfunctioning (all the time) | Detect Varying Cheating/ Malfunctioning (off-peak/ on-peak/ all the time) | Detect Intermittent Cheating/ Malfunctioning (irregular time intervals) | Consider TLs | Highlights |
|---|---|---|---|---|---|---|
| LR-ETDM | MLR | ✓ | ✗ | ✗ | ✗ | Require less metered data |
| CVLR-ETDM | MLR with Categorical Variables | ✓ | ✓ | Can be achieved with minor modification | ✗ | |
| ADF | LP | ✓ | ✗ | ✗ | ✓ | Higher DR |
| Enhanced ADF | LP | ✓ | ✓ | ✓ | ✓ | Higher DR & able to detect intermittent NTLs |

time, the number of FP increases. Therefore, it is important to take into account TLs and measurement noise/error in the design of anomaly detection framework to improve the robustness and accuracy of NTL detection analysis. Besides identifying theft and irregularities in meter readings during specific off-peak/on-peak periods, the proposed LP-based Enhanced ADF can still detect meter irregularities even if there are intermittent NTLs. In other words, Enhanced ADF is not restricted to NTL detection during off-peak/on-peak periods only. The results are detailed in Sections 6.3.2.2 and 6.4.2.2.

## 6.8 Summary of Chapter

In this chapter, the performance and reliability of the proposed LR-based and LP-based frameworks in Chapter 3 and 4, respectively, are evaluated and discussed. Particularly, the discussion focuses mainly on validating the achievement of the objectives of this thesis. Performance comparison studies are conducted to study the strengths and weaknesses of the two proposed anomaly detection frameworks in order to determine which framework to be deployed subject to the availability of data and type of NTL event.

Table 6.21 shows a summary of all the proposed schemes. LR-ETDM and CVLR-ETDM which adopted MLR do not consider TLs in the NTL detection analysis, hence the detection rate is lower as compared to the LP-based ADF and Enhanced ADF in the

University of Malaya

presence of TLs. LP is chosen instead of MLR in ADF and Enhanced ADF because of the non-multicollinearity characteristic of MLR. MLR is unable to estimate the coefficients accurately when multicollinearity is present (Studenmund, 2014). In other words, when the predictors are significantly correlated due to the fact that $c_{t_i} \approx p_{t_i,1} + p_{t_i,2} + \cdots + p_{t_i,N}$, MLR cannot be adopted to solve the LSE in Equations (4.11) and (4.18). The LR-based anomaly detection schemes require less metered data as compared to LP-based ones to detect constant and varying NTL activities. However, to detect more sophisticated and intermittent NTLs such as irregular partial meter bypass, metered data are observed over longer periods as more data samples are required for detection analysis in Enhanced ADF. Therefore, specific detection framework is selected based on the data availability and type of NTL.

University of Malaya

# CHAPTER 7: CONCLUSION

Energy theft is a daunting global problem that results in high utility costs and increased costs to benign paying consumers, as well as a range of safety issues. In recent years, SM and other Internet-based software in SG have increased the chances for energy theft, yet the UPs are still incapable of identifying the sophisticated attacks that target the metering infrastructure. Therefore, anomaly detection framework that identify consumers' energy consumption patterns that are indicative of NTL activities is necessary to thwart electricity pilfering from SM. In this concluding chapter, the key findings of the preceding chapters are summarized and several interesting future research directions are suggested.

## 7.1 Summary of Key Findings

The work in this thesis has achieved the objectives outlined in Section 1.3 by putting forward two anomaly detection frameworks using regression and optimization analyses.

Firstly, a metric known as *anomaly coefficient* is proposed to model the amount of stolen energy at each SM in order to detect the localities of under-reporting and over-reporting by malicious SMs, i.e., a LR-based scheme for Detection of Energy Theft and Defective Smart Meters (LR-ETDM) is put forward in Chapter 3 to detect constant-rate under-reporting and over-reporting by SMs. However, it is shown that varying-rate cheating/malfunctioning in energy reporting might cause some of the fraudulent consumers to escape detection when their cheating behaviors change within the period of observations. To overcome the deficiency of LR-ETDM, Categorical Variable-Enhanced LR-ETDM (CVLR-ETDM) is proposed to resolve the varying-rate cheating/malfunctioning problem. *Categorical variables* are introduced in linear regression to categorize the period of energy fraud and meter irregularities. In addition, another metric referred to as the *detection coefficient* is also introduced to capture the changes of the anomaly coefficients in order to detect the

University of Malaya

period of NTL activities. The simulation and hardware experimentation results show that the suspected consumer can be deduced whether he/she is committing theft either all the time or only during a particular period in a day by investigating the estimated anomaly coefficient and detection coefficient of each consumer.

In the LR-based detection framework, the work in this thesis assumes that power line losses are known, which in practice may be difficult to obtain. In pursuit of higher DR and lower FP, a LP-based anomaly detection framework, known as ADF, is designed in Chapter 4 to take into consideration TLs and calibration error of the equipment for more accurate and efficient anomaly detection. A metric referred to as *loss factor* is introduced to estimate the percentage of TLs in the service area. Furthermore, another variable known as the *error term* is also designed to approximate the random calibration noise/error of the measuring equipment. Then, in order to detect fraudulent consumers' intermittent and more sophisticated malicious behaviors, an Enhanced ADF scheme is put forward. In Enhanced ADF, consumers' reported SM readings are analyzed over a longer period according to specific time slot. As a result of separating consumers' anomaly coefficient evaluation according to time slot, the results indicate that Enhanced ADF is able to detect the localities of malicious events even when there are intermittent cheating and/or faulty equipment, and not restricted to detection during off-peak and on-peak periods only.

A diverse set of NTL attack functions is investigated and generated such that the experiments are closely related to the possible real-world energy fraud/meter irregularities scenarios. From the simulation and hardware experimentation results in Chapter 6, it has been established that ADF outperforms LR-ETDM by a considerable margin in the presence of TLs, i.e., less FP and higher DR in detecting constant under-reporting and over-reporting by SMs, suggesting that the impact caused by TLs and measurement noise at the anomaly detection analysis may be substantial. The detection accuracy might be affected if the

University of Malaya

amounts of both TLs and measurement error are not accounted for in the detection analysis. Similarly, results also suggested that Enhanced ADF outperforms CVLR-ETDM with the presence of TLs, as loss factor and error term capture the percentage of TLs and noise in the system, respectively, thereby improving the detection accuracy. However, sample size versus accuracy trade-off is observed in the Enhanced ADF scheme as it requires more observation data as compared to CVLR-ETDM. Due to this trade-off, it is observed that Enhanced ADF may no longer be beneficial for NTL detection when the data sample size is less than the number of consumers in the service area. In short, LR-based anomaly detection framework is able to identify the positions of energy thieves and faulty SMs without requiring large volume of data samples. On the other hand, LP-based framework is more robust as compared to LR-based because the former is capable of detecting more sophisticated types of energy theft/meter irregularities accurately even in the presence of TLs/calibration error. Consequently, the selection of the specific detection framework is based on the data availability and type of NTL. The results also indicate that the two proposed frameworks are able to realize faster, greater flexibility and improved practicality in the detection of energy theft/meter irregularities based on a small volume of consumers' energy consumption data samples. Moreover, both proposed frameworks can be extended easily to accommodate more consumers for anomaly detection. They are more robust as compared to most existing detection schemes due to the advantages of regression and optimization analyses.

Overall, the studies in Chapter 3 through Chapter 6 have demonstrated how the detection rate of the proposed frameworks is influenced by certain factors (i.e., technical losses and measurement noise of equipment), how categorical variables are able to improve the detection rate of CVLR-ETDM, how to detect intermittent NTL events by separating consumers' anomaly coefficient evaluation according to time slot and how multiple linear

University of Malaya

regression estimation and optimization analyses scale with the number of consumers in the neighborhood.

The research conducted in this thesis are of great practical significance in assisting utility providers to reduce costs incurred due to NTLs and meter irregularities in smart grid environment. No extra hardware costs will incur as utility providers can directly apply the proposed frameworks to detect the localities of defective and compromised smart meters entirely based on the collected energy consumption data. Particularly, all the frameworks proposed in this dissertation involve the study of consumers' energy consumption behavior to detect the amount of stolen/over-reported energy at the smart meters with respect to the discrepancies of meter readings (i.e., energy balance analysis). This in turn reduces the overall operation costs of utility providers and paying prices for consumers. The advantages and limitations of the frameworks will be discussed in the next section.

## 7.2 Advantages and Limitations

The outcome of this research has the following advantages. The proposed frameworks outperform existing work because linear regression and linear programming analyses are:

1. **Protected against contamination attacks**. Unlike most state-of-the-art classification-based detection scheme (i.e., SVM and ELM), polluted dataset and granular changes in data may not affect the detection rate as the proposed frameworks do not require training using historical data.
2. **Robust against non-malicious factors** such as seasonality, appliance and residential changes. The proposed frameworks are free from influences by non-malicious factors as they are not trained by historical data.
3. **Not restricted by the dimension of consumers' power consumption data**. Both proposed anomaly detection frameworks can successfully identify all the malicious

University of Malaya

and defective SMs in the NAN regardless of the dimension of the metered energy consumption data.

4. **Able to reveal the amount of energy theft/loss based on a small volume of consumers' power consumption data samples** regardless of TLs, measurement noise/errors and the type of consumer.
5. **Able to detect the localities of anomalous and compromised smart meters entirely based on collected smart meter readings without incurring extra hardware and software costs**.

Although promising results are attained, the proposed frameworks in this thesis still pose several shortcomings:

1. The proposed anomaly detection frameworks detect NTLs based on the energy balance analysis. Hence, the frameworks are unable to detect energy theft attack that evades the balance check. For instance, an energy thief who compromises a neighbor's SM to ensure that the consumption of at least one of his/her neighbors is over-reported may escape from the anomaly detection (Y. Liu et al., 2018). In such a case, the innocent neighbor would be required to pay for the stolen energy charges.
2. Considering the dynamic pricing (i.e., TOU) in SGs, energy theft is also possible by changing the order of SM readings without altering the average. Energy thieves who change the order of meter readings may escape from detection. In the proposals, a linear relationship is assumed between the dependent variable (i.e., mismatch of meter readings $y_{t_i}$) and the independent variables (i.e., SM readings $p_{t_i,n}$). When the order of meter readings are changed, the relationship between the variables becomes non-linear and hence the detection becomes unpredictable.

University of Malaya

## 7.3 Future Works

All the frameworks proposed in this dissertation involve the study of consumers' energy consumption behavior to detect the amount of stolen/over-reported energy at the SMs with respect to the discrepancies of meter readings. As a result, these frameworks are unable to detect NTL events when fraudulent consumers change the order of the meter readings. In addition, energy theft that evades energy balance check may stay undetected. As a future research direction, techniques such as state estimation, Kalman filter can be applied to detect these cleverly-crafted electricity theft attacks that circumvent detectors.

Furthermore, the work in this thesis assumes that the DER generation measurements recorded by the generation meter in the microgrid are genuine and not manipulated. Therefore, the detection of malicious consumers who over-report the energy they generate for financial gain (i.e., feed-in tariff theft) can be considered as future work.

Last but not least, as discussed in Section 2.2.4.1, most existing energy theft detection schemes require the collection of fine-grained energy consumption data, e.g., consumers' load profiles, which would constitute a potential privacy threat to the consumers (McDaniel & McLaughlin, 2009). Although deploying SG has immense benefits, several privacy concerns arise. As future work, the proposed anomaly detection frameworks can be improved to encrypt consumers' SM meter readings while still being able to identify the locations of malicious and defective SMs.

University of Malaya

## REFERENCES


Abu, H. A., Saharuddin, S., Hussein, Z. F., Malathy, B., Busrah, A. M., & Devaraju, P. (2013). *TNB technical guidebook on grid-interconnection of photovoltaic power generation system to LV and MV networks*. TNB.

Accenture. (2011). *Achieving high performance with theft analytics.* Retrieved from http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Achieving-High-Performance-with-Theft-Analytics.pdf

Ahmad, T., Chen, H., Wang, J., & Guo, Y. (2018). Review of various modeling techniques for the detection of electricity theft in smart grid environment. *Renewable and Sustainable Energy Reviews*, *82*, 2916–2933.

Amin, S., Schwartz, G. A., Cardenas, A. A., & Shankar Sastry, S. (2015). Game-theoretic models of electricity theft detection in smart utility networks. *IEEE Control Systems Magazine*(February), 66–81.

Amral, N., Ozveren, C. S., & King, D. (2007). Short term load forecasting using multiple linear regression. In *2007 42nd International Universities Power Engineering Conference* (pp. 1192–1198).

Artes, M. (1997). Statistical errors. *Medicina Clinica*, *109*(15), 606–607.

Au, M. T., Anthony, T. M., Kamaruddin, N., Verayiah, R., Mustaffa, S. A., & Yusoff, M. (2008). A simplified approach in estimating technical losses in distribution network based on load profile and feeder characteristics. In *2008 IEEE 2nd International Power and Energy Conference* (pp. 1661–1665). Johor Bahru, Malaysia.

Au, M. T., & Tan, C. H. (2013). Energy flow models for the estimation of technical losses in distribution network. *IOP Conference Series: Earth and Environmental Science*, *16*(1), 012035.

Beauty centre caught stealing electricity using remote control switch. (2014, June). *The Star Online*. Retrieved from https://www.thestar.com.my/news/community/2014/06/19/beauty-centre-caught-stealing-electricity-using-remote-control-switch/

Benedict, E. (1992). Losses in electric power systems. *Electrical and Computer Engineering Technical Report. Purdue e-pubs. Purdue University*.


University of Malaya

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. New York, NY, USA: Cambridge University Press.

Buzau, M. M., Tejedor-Aguilera, J., Cruz-Romero, P., & Gomez-Exposito, A. (2018). Detection of non-technical losses using smart meter data and supervised learning. *IEEE Transactions on Smart Grid*, 1–10.

Cardenas, A. A., Amin, S., Schwartz, G., Dong, R., & Sastry, S. (2012). A game theory model for electricity theft detection and privacy-aware control in AMI systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton 2012)* (pp. 1830–1837). Monticello, IL, USA.

Chambers, M., & Dinsmore, T. W. (2014). *Advanced analytics methodologies driving business value with analytics* (1st ed.). Pearson Education, Inc.

Comed. (2017). *Safeguarding data through smarter technology.* Retrieved from https://www.comed.com/SiteCollectionDocuments/SmartEnergy/SmartGridAndDataSecurity.pdf

Commission for Energy Regulation. (2009). *CER Smart Metering Project - Electricity customer behaviour trial, 2009-2010.* Irish Social Science Data Archive. Retrieved from http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

Congres International des Reseaux Electriques de Distribution. (2017). *Reduction of technical and non-technical losses in distribution networks.* Retrieved from www.cired.net/files/download/188

Cuijpers, C., & Koops, B.-J. (2012). Smart metering and privacy in Europe: Lessons from the Dutch case. *European Data Protection: Coming of Age*, 269–293.

Depuru, S. S. S. R., Wang, L., & Devabhaktuni, V. (2011). Smart meters for power grid: Challenges, issues, advantages and status. *Renewable and Sustainable Energy Reviews*, *15*(6), 2736–2742.

Dortolina, C. A., & Nadira, R. (2005). The loss that is unknown is no loss at all: A top-down/bottom-up approach for estimating distribution losses. *IEEE Transactions on Power Systems*, *20*(2), 1119–1125.

Dos Angelos, E. W. S., Saavedra, O. R., Cortés, O. A. C., & de Souza, A. N. (2011).

University of Malaya

Detection and identification of abnormalities in customer consumptions in power distribution systems. *IEEE Transactions on Power Delivery*, *26*(4), 2436–2442.

Electricity theft uncovered at massage parlour, snooker centre. (2016, November). *Borneo Post Online*. Retrieved from http://www.theborneopost.com/2016/11/14/electricity-theft-uncovered-at-massage-parlour-snooker-centre/

Energy Commission. (2015). The national grid: Strengthening Malaysia's framework. *Energy Malaysia*, *6*.

Engel, D. (2013). Privacy and security challenges in the smart grid user domain. In *The first ACM workshop on Information hiding and multimedia security - IH&MMSec '13* (p. 85). Montpellier, France.

Evanczuk, S. (2015). Employing tamper detection and protection. *Electronic Products*. Retrieved from https://www.digikey.com/en/articles/techzone/2015/jun/employing-tamper-detection-and-protection-in-smart-meters

Fang, X., Misra, S., Xue, G., & Yang, D. (2012). Smart grid – The new and improved power grid: A survey. *IEEE Communications Surveys & Tutorials*, *14*(4), 944–980.

Foster, S. (2017). *Non-technical losses: A $96 billion global opportunity for electrical utilities.* Retrieved from http://www.pennenergy.com/articles/pennenergy/2017/11/non-technical-losses-a-96-billion-global-opportunity-for-electrical-utilities.html

Han, W., & Xiao, Y. (2014). NFD : A practical scheme to detect non-technical loss fraud in smart grid. *IEEE ICC 2014-Communication and Information Systems Security Symposium*, 605–609.

Huang, S.-C., Lo, Y.-L., & Lu, C.-N. (2013). Non-technical loss detection using state estimation and analysis of variance. *IEEE Transactions on Power Systems*, *28*(3), 2959–2966.

IEEE Power & Energy Society. (1992). *Distribution test feeders.* Retrieved from http://sites.ieee.org/pes-testfeeders/resources/

Jiang, R., Lu, R., Wang, Y., Luo, J., Shen, C., & Shen, X. (2014). Energy-theft detection issues for advanced metering infrastructure in smart grid. *Tsinghua Science and Technology*, *19*(2), 105–120.

University of Malaya

Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2016). Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, *12*(3), 1005 – 1016.

Jokar, P., Arianpoo, N., & Leung, V. C. (2016). Electricity theft detection in AMI using customers' consumption patterns. *IEEE Transactions on Smart Grid*, *7*(1), 216–226.

Jovanovic, P., & Neves, S. (2015a). Dumb crypto in smart grids: Practical cryptanalysis of the open smart grid protocol. *IACR Cryptology ePrint Archive*, *2015*, 428.

Jovanovic, P., & Neves, S. (2015b). Practical cryptanalysis of the open smart grid protocol. In *International Workshop on Fast Software Encryption* (pp. 297–316). Springer.

Karaim, R. (2015). *Power theft: Co-ops use high-tech and low-tech strategies to detect theft.* RE Magazine. Retrieved from http://remagazine.coop/power-theft/

Khoo, B., & Cheng, Y. (2011). Using RFID for anti-theft in a chinese electrical supply company : A cost-benefit analysis. In *Wireless Telecommunications Symposium (WTS 2011)* (pp. 1–6). New York City, NY, USA.

Koberstein, A. (2008). Progress in the dual simplex algorithm for solving large scale LP problems: Techniques for a fast and stable implementation. *Computational Optimization and Applications*, *41*(2), 185–204.

Krishna, V. B., Iyer, R. K., & Sanders, W. H. (2016). ARIMA-based modeling and validation of consumption readings in power grids. In *10th International Conference on Critical Information Infrastructures Security* (Vol. 9578, pp. 199–210). Berlin, Germany.

Krishna, V. B., Lee, K., Weaver, G. A., Iyer, R. K., & Sanders, W. H. (2016). F-DETA: A framework for detecting electricity theft attacks in smart grids. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (pp. 407–418). Toulouse, France.

Krishnaswamy, V. (1999). *Non-payment in the electricity sector in Eastern Europe and the Former Soviet Union (English).* Washington, D.C.: The World Bank. Retrieved from http://documents.worldbank.org/curated/en/221271468758976599/Non-payment-in-the-electricity-sector-in-Eastern-Europe-and-the-Former-Soviet-Union

University of Malaya

Li, F., Qiao, W., Sun, H., Wan, H., Wang, J., Xia, Y., ... Zhang, P. (2010). Smart transmission grid: Vision and framework. *IEEE Transactions on Smart Grid*, *1*(2), 168–177.

Liu, J., Xiao, Y., & Gao, J. (2014). Achieving accountability in smart grid. *IEEE Systems Journal*, *8*(2), 493–508.

Liu, Y., Zhou, Y., & Hu, S. (2018). Combating coordinated pricing cyberattack and energy theft in smart home cyber-physical systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *37*(3), 573–586.

Marris, E. (2008). Energy: Upgrading the grid. *Nature*, *454*(7204), 570–573.

Mashima, D., & Cárdenas, A. a. (2012). Evaluating electricity theft detectors in smart grid networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7462 LNCS, pp. 210–229).

MathWorks. (2017). *Matlab official website.* Retrieved from https://www.mathworks.com/products/matlab.html

McDaniel, P., & McLaughlin, S. (2009). Security and privacy challenges in the smart grid. *IEEE Security and Privacy*, *7*(3), 75–77.

McLaughlin, S., Holbert, B., Fawaz, A., Berthier, R., & Zonouz, S. (2013). A multi-sensor energy theft detection framework for advanced metering infrastructures. *IEEE Journal on Selected Areas in Communications*, *31*(7), 1319–1330.

McLaughlin, S., Podkuiko, D., & McDaniel, P. (2010). Energy theft in the advanced metering infrastructure. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6027 LNCS, pp. 176–187).

Messinis, G. M., & Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods. *Electric Power Systems Research*, *158*, 250–266.

Meuse, M. (2016, November). BC Hydro uses new technology to stop theft, spot grow-ops. *CBC News*. Retrieved from http://www.cbc.ca/news/canada/british-columbia/hydro-grid-meters-1.3837496

University of Malaya

Millard, R., & Emmerton, M. (2009). Non technical losses – how do other countries tackle the problem? In *22nd AMEU Technical Convention* (pp. 1–14). Pretoria.

Muniz, C., Figueiredo, K., Vellasco, M., Chavez, G., & Pacheco, M. (2009). Irregularity detection on low tension electric installations by neural network ensembles. In *International Joint Conference on Neural Networks* (pp. 2176–2182). Atlanta, GA.

Murrill, B. J., Liu, E. C., & Thompson II, R. M. (2012, February). *Smart meter data: Privacy and cybersecurity.* Congressional Research Service. Retrieved from https://fas.org/sgp/crs/misc/R42338.pdf

Nadira, R., Benchluch, S., & Dortolina, C. (2003). A novel approach to computing distribution losses. In *2003 IEEE PES Transmission and Distribution Conference and Exposition* (Vol. 2, pp. 3–7). Dallas, USA.

Nagi, J. (2009). *An intelligent system for detection of non-technical losses in Tenaga Nasional Berhad (TNB) Malaysia low voltage distribution network* (Master's thesis, Universiti Tenaga Nasional). Retrieved from http://people.idsia.ch/~nagi/thesis/mee_thesis.pdf

Nagi, J., Mohammad, A., Yap, K., Tiong, S., & Ahmed, S. (2008). Non-technical loss analysis for detection of electricity theft using support vector machines. In *2008 IEEE 2nd International Power and Energy Conference* (pp. 907–912). Johor Bahru, Malaysia.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohamad, M. (2010). Nontechnical loss detection for metered customers in power utility using support vector machines. *IEEE Transactions on Power Delivery*, *25*(2), 1162–1171.

Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., & Mohammad, a. M. (2008). Detection of abnormalities and electricity theft using genetic support vector machines. In *TENCON 2008 - 2008 IEEE Region 10 Conference* (pp. 1–6). Hyderabad, India.

Navani, J. P., Sharma, N. K., & Sapra, S. (2012). Technical and non-technical losses in power system and its economic consequence in Indian economy. *International Journal of Electronics and Computer Science Engineering*, *1*(2), 757–761.

Nikovski, D. N., Wang, Z., Esenther, A., Sun, H., Sugiura, K., Muso, T., & Tsuru, K. (2013). Smart meter data analysis for power theft detection. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 379–389). Springer Berlin Heidelberg.

University of Malaya

Nizar, A. H., Dong, Z. Y., Jalaluddin, M., & Raffles, M. J. (2006). Load profiling method in detecting non-technical loss activities in a power utility. In *2006 IEEE International Power and Energy Conference* (pp. 82–87).

Nizar, A. H., Dong, Z. Y., & Wang, Y. (2008). Power utility nontechnical loss analysis with extreme learning machine method. *IEEE Transactions on Power Systems*, *23*(3), 946–955.

Nizar, A. H., Dong, Z. Y., Zhao, J. H., & Zhang, P. (2007). A data mining based NTL analysis method. In *2007 IEEE Power Engineering Society General Meeting* (pp. 1–8). Tampa, FL, USA.

Nizar, A. H., Zhao, J. H., & Dong, Z. Y. (2006). Customer information system data pre-processing with feature selection techniques for non-technical losses prediction in an electricity market. In *International Conference on Power System Technology* (pp. 1–7). Chongqing, China.

Northeast Group. (2017, May). $96 billion is lost every year to electricity theft. *PR Newswire*. Retrieved from http://www.prnewswire.com/news-releases/96-billion-is-lost-every-year-to-electricity-theft-300453411.html

Oliveira, M. E., & Padilha-Feltrin, A. (2009). A top-down approach for distribution loss evaluation. *IEEE Transactions on Power Delivery*, *24*(4), 2117–2124.

Omron. (2017). *NJ-series machine automation controller database connection CPU unit.* Retrieved from https://www.valin.com/sites/default/files/asset/document/Omron-NJ-Database-CPU-Brochure.pdf

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (Vol. 3). New York Harcourt Brace College Publishers.

Pedro, A. (2009). Reducing technical and non-technical losses in the power sector. *World Bank Group Energy Sector Strategy*, 35. Retrieved from https://openknowledge.worldbank.org/handle/10986/20786

Rashed Mohassel, R., Fung, A., Mohammadi, F., & Raahemifar, K. (2014). A survey on advanced metering infrastructure. *International Journal of Electrical Power and Energy Systems*, *63*, 473–484.

University of Malaya

Refou, O., Alsafasfeh, Q., & Alsoud, M. (2015). Evaluation of electric energy losses in southern governorates of Jordan distribution electric system. *International Journal of Energy Engineering*, *5*(2), 25–33.

Rodriguez, G. (2013). Lecture notes on generalized linear models. *Princeton Statistics*. Retrieved from http://data.princeton.edu/wws509/notes/

Sahoo, S., Nikovski, D., Muso, T., & Tsuru, K. (2015). Electricity theft detection using smart meter data. In *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)* (pp. 1–5). Columbia, USA.

Salinas, S., Li, M., & Li, P. (2013). Privacy-preserving energy theft detection in smart grids: A P2P computing approach. *IEEE Journal On Selected Area In Communications/Supplement*, *31*(9), 257–267.

Salinas, S. A., & Li, P. (2015). Privacy-preserving energy theft detection in microgrids: A state estimation approach. *IEEE Transactions on Power Systems*, 1–12.

Sankar, L., Raj Rajagopalan, S., Mohajer, S., & Vincent Poor, H. (2013). Smart meter privacy: A theoretical framework. *IEEE Transactions on Smart Grid*, *4*(2), 837–846.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: Part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt international*, *107*(44), 776–782.

Selvapriya, C. (2014). Competent approach for inspecting electricity theft. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(3), 1763–1766.

Siddiqui, O. (2008). *The green grid: Energy savings and carbon emissions reductions enabled by a smart grid.* Electric Power Research Institute. Retrieved from https://www.smartgrid.gov/files/The_Green_Grid_Energy_Savings_Carbon_Emission_Reduction_En_200812.pdf

Skrivanek, S. (2009). *The use of dummy variables in regression analysis.* More Steam, LLC. Retrieved from https://www.moresteam.com/whitepapers/download/dummy-variables.pdf

University of Malaya

Smith, T. B. (2004). Electricity theft: A comparative analysis. *Energy Policy*, *32*, 2067–2076.

Spirić, J. V., Stanković, S. S., Dočić, M. B., & Popović, T. D. (2014). Using the rough set theory to detect fraud committed by electricity customers. *International Journal of Electrical Power and Energy Systems*, *62*, 727–734.

Starkweather, J. (1997). *Categorical variables in regression: Implementation and interpretation.* Retrieved from https://it.unt.edu/sites/default/files/categoricalregression_jds_june2010.pdf

Studenmund, A. H. (2014). *Using Econometrics : A practical guide* (6th ed.). Pearson Education Limited.

Suriyamongkol, D. (2002). *Non-Technical losses in electrical power systems* (Master's thesis, Ohio University). Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1175007802

Tellbach, D., & Li, Y.-F. (2018). Cyber-attacks on smart meters in household nanogrid: Modeling, simulation and analysis. *Energies*, *11*(2), 316.

Telles Esteves, G. R., Cyrino Oliveira, F. L., Antunes, C. H., & Souza, R. C. (2016). An overview of electricity prepayment experiences and the Brazilian new regulatory framework. *Renewable and Sustainable Energy Reviews*, *54*, 704–722.

Tenaga Nasional Berhad. (2006). *Tenaga Nasional Berhad annual report 2006.* TNB. Retrieved from https://www.tnb.com.my/assets/annual_report/AR06.pdf

Tenaga Nasional Berhad. (2017). *Integrated annual report (IAR) 2017.* TNB. Retrieved from https://www.tnb.com.my/assets/annual_report/TNB_Annual_Report_2017.pdf

Tenaga Nasional Berhad. (2018). *Meter replacement (due to irregularity).* TNB. Retrieved from https://www.tnb.com.my/faq/meter-replacement-due-to-irregularity/

Tetko, I. V., Livingstone, D. J., & Luik, A. I. (1995). Neural network studies. 1. comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, *35*(5), 826–833.

University of Malaya

U.S. Department of Energy. (2008, February). *The NETL modern grid strategy powering our 21st-century economy: Advanced metering infrastructure.* National Energy Technology Laboratory. Retrieved from https://www.smartgrid.gov/document/netl_modern_grid_strategy_powering_our_21st_century_economy_advanced_metering_infrastructur

U.S. Department of Energy. (2010, October). *Communications requirements of smart grid technologies.* National Energy Technology Laboratory. Retrieved from https://www.energy.gov/sites/prod/files/gcprod/documents/Smart_Grid_Communications_Requirements_Report_10-05-2010.pdf

Viegas, J. L., Esteves, P. R., Melício, R., Mendes, V. M., & Vieira, S. M. (2017). Solutions for detection of non-technical losses in the electricity grid: A review. *Renewable and Sustainable Energy Reviews*, *80*, 1256–1268.

Villar-Rodriguez, E., Del Ser, J., Oregi, I., Bilbao, M. N., & Gil-Lopez, S. (2017). Detection of non-technical losses in smart meter data based on load curve profiling and time series analysis. *Energy*, *137*, 118–128.

Vincenzo, Giordano and Georgios, Papaefthymiou. (2015, December). *Identifying energy efficiency improvements and saving potential in energy networks, including analysis of the value of demand response.* Tractebel Engineering. Retrieved from https://ec.europa.eu/energy/sites/ener/files/documents/GRIDEE_4NT_364174_000_01_TOTALDOC%20-%2018-1-2016.pdf

Wafi, A., Aziz, A., Rahim, N., Amirhussain, A. H., & Norddin, N. B. (2013). Intelligent Tenaga Nasional Berhad (TNB) single phase power supply cut-off. *Journal of Telecommunication, Electronic and Computer Engineering*, *9*(2), 55–58.

Wang, W., & Lu, Z. (2013). Cyber security in the smart grid: Survey and challenges. *Computer Networks*, *57*(5), 1344–1371.

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 1-1.

Wu, S. (2013). A review on coarse warranty data and analysis. *Reliability Engineering and System Safety*, *114*, 1–11.

Xiao, Z., Xiao, Y., & Du, D. H. C. (2013). Exploring malicious meter inspection in

University of Malaya

neighborhood area smart grids. *IEEE Transactions on Smart Grid*, *4*(1), 214–226.

Xu, V. Z. (2015). *A design of theft detection framework for smart grid network* (Master's thesis, University of Waterloo). Retrieved from http://hdl.handle.net/10012/9837

Yan, Y., Qian, Y., Sharif, H., & Tipper, D. (2013). A survey on smart grid communication infrastructures: Motivations, requirements and challenges. *IEEE Communications Surveys and Tutorials*, *15*(1), 5–20.

Yip, S. C., Tan, W. N., Tan, C., Gan, M. T., & Wong, K. S. (2018). An anomaly detection framework for identifying energy theft and defective meters in smart grids. *International Journal of Electrical Power and Energy Systems*, *101*, 189–203.

Yip, S. C., Wong, K. S., Hew, W. P., Gan, M. T., Phan, R. C., & Tan, S. W. (2017). Detection of energy theft and defective smart meters in smart grids using linear regression. *International Journal of Electrical Power and Energy Systems*, *91*, 230–240.

University of Malaya

## LIST OF PUBLICATIONS AND PAPERS PRESENTED

The following is the list of submitted / accepted journal articles and peer-viewed conference papers related to this study.

**Journals:**

[1] **Yip, S. C.**, Wong, K. S., Hew, W. P., Gan, M. T., Phan, R. C. W., & Tan, S. W. (2017). Detection of energy theft and defective smart meters in smart grids using linear regression. *International Journal of Electrical Power and Energy Systems*, *91*, 230–240. (Impact factor 2018: 3.289).

[2] **Yip, S. C.**, Tan, W. N., Tan, C. K., Gan, M. T., & Wong, K. S. (2018). An anomaly detection framework for identifying energy theft and defective meters in smart grids. *International Journal of Electrical Power and Energy Systems*, *101*, 189–203. (Impact factor 2018: 3.289).

**International Peer-Reviewed Conferences:**

[1] **Yip, S. C.**, Tan, C., Tan, W. N., Gan, M. T., & Abu Bakar, A.-H. (2017). Energy theft and defective meters detection in AMI using linear regression. In 2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe) (pp. 1–6). Milan, Italy.

[2] **Yip, S. C.**, Tan, C., Tan, W. N., Gan, M. T., Wong, K., & Phan, R. C. (2018). Detection of Energy Theft and Metering Defects in Advanced Metering Infrastructure Using Analytics. In 2018 International Conference on Smart Grid and Clean Energy Technologies (pp. 1–8). Kajang, Malaysia.

University of Malaya