Detection of Energy Theft and Metering Defects in Advanced Metering Infrastructure Using Analytics

Sook-Chin Yip*†, ChiaKwang Tan*, Wooi-Nee Tan†, Ming-Tao Gan†, KokSheik Wong‡, Raphael C.-W. Phan†
*UM Power Energy Dedicated Advanced Center (UMPEDAC), University of Malaya, 50603 Kuala Lumpur, MALAYSIA.
†Faculty of Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, MALAYSIA.
‡School of Information Technology, Monash University Malaysia, MALAYSIA.
Email: scyip@mmu.edu.my, cktan@um.edu.my, wntan@mmu.edu.my, mtgan@mmu.edu.my,
wong.koksheik@monash.edu, raphael@mmu.edu.my

Abstract — **Non-technical losses including electricity theft and anomalies in meter readings are estimated to cost the utility providers tremendous losses of approximately $96 billion per annum. The adoption of smart meter has encouraged utility providers to use analytics to identify theft. To curb non-technical losses, they are increasingly leveraging on real-time smart metering and analytics to identify energy theft and irregularities in meter readings. We have previously put forward linear regression-based and linear programming-based anomaly detection frameworks to study consumers' energy consumption behavior for detecting the localities of metering defects as well as energy thefts. In this work, we design and construct an advanced metering infrastructure test rig in the laboratory to perform comparison studies on our previously proposed anomaly detection frameworks in smart grid environment. Results from both test rig and simulations show that linear regression-based anomaly detection framework is able to identify the positions of energy thieves and faulty smart meters without requiring large volume of data samples. However, linear programming-based framework is more robust as compared to linear regression-based because the former is capable of detecting more sophisticated types of energy theft/meter irregularities accurately even in the presence of technical losses/calibration errors.**

Keywords— non-technical losses detection; technical losses; AMI; linear regression; linear programming; analytics;

I. Introduction

Electricity theft, which is also known as non-technical losses (NTLs) is crippling utility providers (UPs) around the world. NTLs not only result in costly government subsidies and higher paying prices for consumers but also public safety concerns with life-threatening illegal power connections. According to Northeast Group, NTLs cost a staggering $96 billion per year globally [1]. The World Bank also reported that almost half of the generated energy in developing countries is lost through energy theft [2].

The deployment of advanced metering infrastructure (AMI) in smart grids (SGs) has become significantly imperative for combating NTLs as well as providing higher reliability, better power quality, lower utility costs and more accurate billing. Despite these technical and societal benefits, AMI is still vulnerable to more sophisticated types of malicious attack. Specifically, smart meter (SM) enables functionalities such as automatic transmission of metering data and remote update of firmware. However, these features also create a back door for malicious adversaries. For instance, an adversary can gain the root access of the SM and falsify the meter readings to reduce billings [3]. To address NTLs, UPs are progressively leveraging on real-time smart metering in AMI and analytics to identify areas of concentrated energy theft/meter irregularities.

The existing classification-based NTLs detection techniques differentiate abnormal energy usage patterns from all energy usage patterns based on a testing dataset containing samples of both the normal and anomalous classes. Historical consumption data, a data mining method together with support vector machine (SVM) classifier were used to detect abnormal behaviors in [4]. In [5], Jindal *et al.* proposed a decision tree and SVM-based data analytics to detect and locate real-time electricity theft at every level in power transmission and distribution system. Meanwhile, Krishna *et al.* [6] proposed an anomaly detection method which combines principal component analysis and density-based spatial clustering of applications with noise to verify the integrity of the SM measurements. In another work [7], they adopted the Kullback-Leibler divergence to identify cleverly-crafted energy fraud attacks which circumvent detectors. Nonetheless, one of the drawbacks of these detection methods is the vulnerability to contamination attacks. Specifically, an adversary can mislead the learning machine to classify an anomalous pattern as a normal one through data pollution and granular changes in data. Besides that, some of these SVM-based detection schemes typically require long-term measurements and monitoring before NTLs detection can be performed precisely. The large sample size requirement normally leads to longer detection delay [8]. Another challenging issue that affects the performance of some classification-based methods is the fact that several non-malicious factors such as change of appliances, seasonality, etc. can alter the consumption pattern. If these factors are not properly dealt with, they might result in high false positive rate.

On the other hand, Salinas *et al.* [9] designed a lower-upper decomposition (LUD)-based algorithm to solve a linear system of equations (LSE) for assessing consumers' honesty coefficients while protecting their privacy. Nevertheless, their proposal is restricted by the dimension of the metered data. The data must be in square matrix due to the characteristic of

978-1-5386-6410-0/18/$31.00 ©2018 IEEE

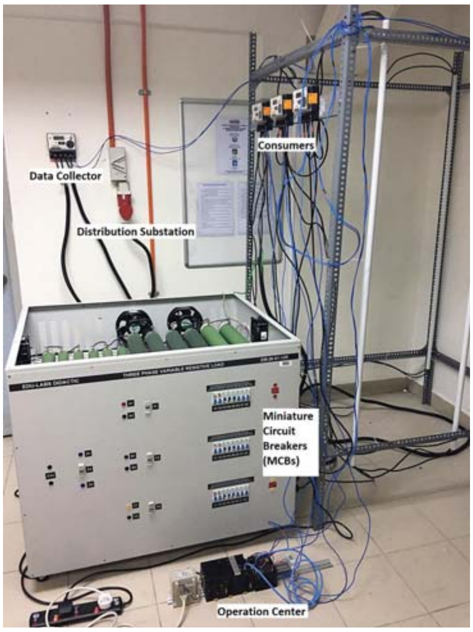


Fig. 1. The hardware experimentation of the AMI test rig in the laboratory [11].

LUD. To meet the dimension requirement, the authors have to change the time granularity of the SM. Nevertheless, it might be impractical to reduce the time granularity indefinitely due to the limited memory size of SM.

To address some of the aforementioned limitations of existing classification-based detection schemes, linear regression-based [10] and linear programming-based [11] anomaly detection frameworks for identifying energy thefts and defective meters are proposed. The key ideas of the proposed frameworks are to adopt *multiple linear regression* (MLR) and *linear programming* (LP) for evaluating consumers' honesty level in energy reporting based on the recorded energy consumption data collected from AMI. The main contributions to this paper are as follows:

1) Construct an AMI test rig in the laboratory to assess the performance of our previously proposed anomaly detection frameworks in smart grid environment, and;
2) Conduct performance comparison studies to study the strengths and weaknesses of the two proposed anomaly detection frameworks.

II. AMI Test Rig Hardware Experimentation

In this work, an AMI test rig, consisting of three consumers, an operation center (i.e., Omron NJ101-1020 machine automation controller [12]) and a distribution substation equipped with data collector is built in the laboratory as shown in Fig. 1 to perform comparison studies on our previously proposed anomaly detection frameworks in smart grid environment. As shown in Fig. 1, a master SM, which is also known as the *data collector* is endowed to measure the aggregate energy consumption supplied to the Neighborhood Area Network (NAN) while three single-phase smart meters (SMs) are used to record the energy consumption data of each consumer. The data collector and all consumers' SMs are configured to record their power consumption readings at half-hourly interval. The `rand` function packaged in Matlab R2014b is used to generate random load demand for each consumer at each 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. The single-line diagram and hardware installation of the test rig are detailed in our previous work [11], [13].

TABLE I
Description of a_n

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

III. Anomaly Detection Frameworks

A. Linear Regression-Based Detection Framework for Energy Theft and Defective Smart Meters

We have previously put forward a **L**inear **R**egression-based Detection scheme for **E**nergy **T**heft and **D**efective Smart **M**eters (LR-ETDM) to identify potential energy thieves and metering defects [10], [13]. In the proposal, we first assumed that the anomaly coefficients are constant. In other words, fraudulent consumers steal energy consistently while there could be faulty SMs that are out of order all the time. Consider a NAN consisting of N consumers. The reading of each SM is recorded at time stamp $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$, for the consumption period $t_i - t_{i-1}$. The parameters in our proposed LR-based framework are defined as follows:

$p_{t_i,n}$ = energy consumption (in kWh) recorded by consumer n at time interval t_i,
a_n = anomaly coefficient for each consumer n, and;
y_{t_i} = discrepancy in meter reading at time interval t_i.

In the event of any under-reporting/over-reporting by SMs, the following can be formulated:

$$a_1 p_{t_i,1} + a_2 p_{t_i,2} + \cdots + a_n p_{t_i,n} = y_{t_i}, t_i \in \mathbf{T}. \tag{1}$$

Our goal is to solve all $a_n, n \in \mathbf{N} = \{1, 2, \cdots, N\}$ from (1) using MLR to validate the anomalous behavior of each consumer or reliability of SM installed in each household. There are three possibilities as shown in Table I.

To detect NTLs which occur only during a certain time in a day, we designed a **C**ategorical **V**ariable-enhanced **LR-ETDM** (CVLR-ETDM). A metric known as the detection coefficient, β_n is introduced in the framework to indicate whether consumer n cheats inconsistently in a day. Categorical variable, x_n is also incorporated into MLR to categorize the period of metering defect or energy theft.

The period of metering defect or energy theft (i.e., off-peak/on-peak/all the time) can be determined by solving for a_n and β_n in

$$a_1 p_{t_i,1} + \cdots + a_n p_{t_i,n} + \beta_1 p_{t_i,1} x_1 + \cdots + \beta_n p_{t_i,n} x_n = y_{t_i}, t_i \in \mathbf{T}. \tag{2}$$

TABLE II
BEHAVIORAL CHARACTERIZATION BASED ON a_n, β_n AND $(a_n + \beta_n)$

Scenario	a_n	β_n	$(a_n + \beta_n)$	Behavior
1	= 0	= 0	= 0	Honest all the time
2	> 0	= 0	> 0	Under-report all the time
3	< 0	= 0	< 0	Over-report all the time
4	= 0	> 0	> 0	Under-report during on-peak
5	= 0	< 0	< 0	Over-report during on-peak
6	> 0	$-a_n$	= 0	Under-report during off-peak
7	$-\beta_n$	> 0	= 0	Over-report during off-peak

In (2), a_n itself denotes the anomaly coefficient of consumer n during *off-peak* period while $(a_n + \beta_n)$ represents the anomaly coefficient of consumer n during *on-peak* period. By investigating the estimated a_n, β_n and the corresponding p-values, we can deduce whether anomalous behavior occurs only during a particular period in a day or all the time as described in Table II. The detailed framework can be found in our earlier work [10].

B. Linear Programming-Based Detection Framework for Energy Theft and Defective Smart Meters

In our first proposed LR-based detection framework, we assume that power line losses are known, which in practice may be difficult to obtain. Thus, we design another LP-based **A**nomaly **D**etection **F**ramework (ADF) to take technical losses (TLs) and calibration errors into consideration for more accurate anomaly detection [11]. A metric referred to as loss factor, l_{t_i} is introduced to (1) for estimating the percentage of TLs in the service area. Another variable known as the error term, E_{t_i} is also introduced to approximate the random calibration noise/error of the equipment. Here, E_{t_i} is expressed as the difference between two non-negative variables, $(E^+)_{t_i}$ and $(E^-)_{t_i}$ to capture the positive and negative measurement errors, respectively. Let $E_{t_i} = (E^+)_{t_i} - (E^-)_{t_i}$, the proposed LP problem is formulated as follows:

$$\text{minimize } f = \sum_{i=1}^{T} \left(\left(E^+\right)_{t_i} + \left(E^-\right)_{t_i} \right)$$

subject to

$$\sum_{n=1}^{N} a_n p_{t_i,n} + l_{t_i} c_{t_i} + (E^+)_{t_i} - (E^-)_{t_i} = y_{t_i}, \forall t_i \in \mathbf{T}, \tag{3}$$

$$(E^+)_{t_i}, (E^-)_{t_i} \geq 0, \forall t_i \in \mathbf{T}, \tag{4}$$

$$a_n \text{ unrestricted}, \forall n \in \mathbf{N}, \tag{5}$$

$$l_{min} \leq l_{t_i} \leq l_{max}, \forall t_i \in \mathbf{T}. \tag{6}$$

A lower f is preferred for higher detection accuracy in pinpointing the potential faulty SMs and fraudulent consumers. According to [14], the average TLs of low voltage (LV) network in Malaysia was reported to range from 0.59% to 3.23%, subject to the capacity and loading of the network. Therefore, we consider that there are 3%-5% of TLs in the service area (i.e., $l_{min} = 0.03$, $l_{max} = 0.05$), as captured in the constraint expressed as (6) to show the practicability of the proposed framework in the presence of TLs.

Then, in order to detect fraudulent consumers' *varying* cheating behaviors, we put forward an **Enhanced ADF** scheme. In Enhanced ADF scheme, the metered energy consumption data are observed over an extended period of time (i.e., at least N days, where N is the number of consumers in the service area) to improve the accuracy of anomaly detection.

Suppose that the reading of each SM is sampled over T time intervals every day for a period of D days. The parameters in the Enhanced ADF scheme are defined as follows:

$p^d_{t_i,n}$ = energy consumption (in kWh) recorded by consumer n at time interval $t_i \in \mathbf{T}$ on day $d \in \mathbf{D}$,
$a_{t_i,n}$ = anomaly coefficient for each consumer n at time interval $t_i \in \mathbf{T}$,
$c^d_{t_i}$ = aggregated energy supplied by the UPs at time interval $t_i \in \mathbf{T}$ on day $d \in \mathbf{D}$,
$l^d_{t_i}$ = loss factor at time interval $t_i \in \mathbf{T}$ on day $d \in \mathbf{D}$, and;
$y^d_{t_i}$ = discrepancy in meter reading at time interval $t_i \in \mathbf{T}$ on day $d \in \mathbf{D}$.

Therefore, the varying anomaly coefficients can be determined by solving the following LP problem:

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

$$\text{minimize } f = \sum_{d=1}^{D} \left(\left(E^+\right)^d_{t_i} + \left(E^-\right)^d_{t_i} \right)$$

subject to

$$\sum_{n=1}^{N} a_{t_i,n} p^d_{t_i,n} + l^d_{t_i} c^d_{t_i} + (E^+)^d_{t_i} - (E^-)^d_{t_i} = y^d_{t_i}, \forall d \in \mathbf{D}, \tag{7}$$

$$(E^+)^d_{t_i}, (E^-)^d_{t_i} \geq 0, \forall d \in \mathbf{D}, \tag{8}$$

$$a_{t_i,n} \text{ unrestricted}, \forall n \in \mathbf{N}, \tag{9}$$

$$l_{min} \leq l^d_{t_i} \leq l_{max}, \forall d \in \mathbf{D}, \tag{10}$$

where $E^d_{t_i}$ represents the error term at time slot $t_i \in \mathbf{T} = \{t_1, t_2, \cdots, t_T\}$ on day $d \in \mathbf{D} = \{1, 2, \cdots, D\}$. As discussed earlier, $E^d_{t_i}$ is separated into $(E^+)^d_{t_i}$ and $(E^-)^d_{t_i}$ to capture the positive and negative measurement errors.

A summary consisting of comparison of techniques as well as the strengths and weaknesses of our proposed frameworks discussed earlier is reported in Section IV-C.

IV. RESULTS AND DISCUSSIONS

In this section, we present the MLR and optimization analysis to assess the performance of our proposed anomaly detection frameworks. Since SGs are not fully implemented in Malaysia yet, real-world SGs energy theft samples are non-existent. Therefore, the smart energy data from the AMI hardware experimentation and Irish Smart Energy Trial [15] are used in our study.

TABLE III
POSSIBLE STATES OF THE SMART METERS

State	State Function	Description
s_1	$p^d_{t_i,n} = 1 \times p^{d*}_{t_i,n}$	SM is neither compromised nor faulty
s_2	$p^d_{t_i,n} = \nu p^{d*}_{t_i,n}; \nu \in (0, 0.95) \cup (1.05, 2.5]$	SM is compromised/faulty all the time
s_3	$p^d_{t_i,n} = \delta_{t_i} p^{d*}_{t_i,n};$ $\delta_{t_i} = \begin{cases} \nu & \text{if } start \leq t_i \leq end \\ 1 & \text{otherwise} \end{cases}$ where ν is as defined in s_2 above;	SM is compromised/faulty only during a certain period in a day

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

In the attack model, the SMs are assumed to be in one of the three states, namely, honest, compromised or out of order. Suppose the SMs in the test rig and [15] are configured to record benign half-hourly meter readings $\mathbf{P^{d*}_n} = \{p^{1*}_{t_1,n}, p^{1*}_{t_2,n}, \cdots, p^{d*}_{t_i,n}\}$ for time intervals $t_i = t_1, t_2, \cdots, t_{48}$ and days $d = 1, 2, \cdots, D$. Table III summarizes the possible states of the n-th SM, $n \in \mathbf{N} = \{1, 2, \cdots, N\}$. Let $p^d_{t_i,n}$ denotes the energy usage recorded by the n-th SM after the application of one of the state functions in Table III. In this work, we assume that either the faulty SM *always* over-reports what was consumed or the dishonest consumer under-reports his/her energy reporting by more than 5%. In such a case, any consumers with anomaly coefficients in $[-0.05, 0.05]$ are classified as honest. That is, a_n, $(a_n + \beta_n)$ or $a_{t_i,n} \approx 0$.

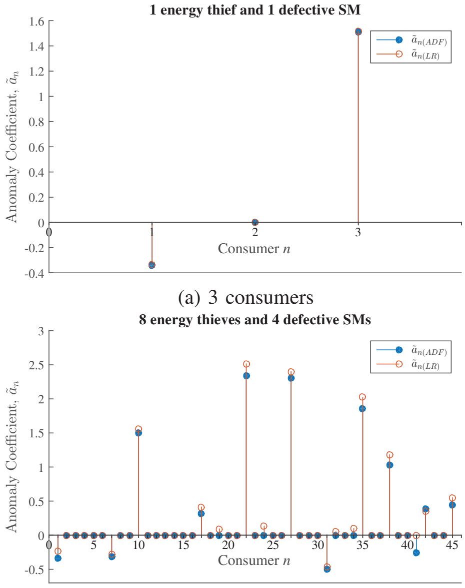
A. Constant Anomaly Coefficients

Here, we compare the performance of LR-ETDM and ADF where the dishonest consumers always under-report their energy consumptions (i.e., case s_2 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. The constant under-reporting/over-reporting scenario for the size of three consumers of the test rig is set up as shown in Table IV. 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 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.

Fig. 2 depicts the values of $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ when a_n is constant. It can be observed from Table IV and Fig. 2(a) that the first SM is faulty 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 suggested that both $\tilde{a}_{n(LR)}$ and $\tilde{a}_{n(ADF)}$ show similar results when the service area is small and the amount of TLs is negligible.

To study how the proposed frameworks scale with the number of consumers in the event of TLs and measurement noise being present, we consider a NAN of 45 consumers. The simulation for the case of 45 consumers is setup in a similar mean as shown in Table V. Four-day half-hourly smart energy data (i.e., 192 data points) from the Irish Smart Energy [15] 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, i.e.,



(a) 3 consumers

(b) 45 consumers

Fig. 2. Values of $\tilde{a}_n$ obtained by LR-ETDM and ADF when a_n is constant

$$c^{d}_{t_i} \leftarrow \left(\sum_{n=1}^{N} p^{d*}_{t_i,n} \right) \div \mu, \mu \in [0.95, 0.97], \forall t_i \in \mathbf{T}, \forall d \in \mathbf{D}. \tag{11}$$

Note that the state-of-the-art SMs measure very accurate measurements, where the errors are usually modeled by white uncorrelated noise with standard deviation and zero mean [16]. Thus, to further validate our proposed frameworks, the measurement errors are considered as below:

$$y^d_{t_i} = c^d_{t_i} - \sum_{n=1}^{N} p^d_{t_i,n} + e^d_{t_i}, \tag{12}$$

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

As shown in Table V and Fig. 2(b), LR-ETDM tends to produce inaccurate anomaly coefficient vector in the presence of TLs and measurement noise. Specifically, some of 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 of 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 $\frac{1}{1+a}$ as shown in Tables IV and V.

TABLE IV

COMPARISON AMONG CONSTANT a_n, $\tilde{a}_{n(LR)}$ AND $\tilde{a}_{n(ADF)}$ OBTAINED FROM HARDWARE EXPERIMENTATION FOR THE SIZE OF 3 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 50%	-0.3333	1.5	-0.3366	1.51	-0.3386	1.51
2	Honest	0	1	$0.0018 \approx 0$	1	0	1
3	Under-report by 60%	1.5	0.4	1.5167	0.40	1.5119	0.40

TABLE V

COMPARISON AMONG CONSTANT a_n, $\tilde{a}_{n(LR)}$ AND $\tilde{a}_{n(ADF)}$ OBTAINED FROM [15] FOR THE 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	1.5	-0.2750	1.38	-0.3144	1.46
10	Under-report by 60%	1.5	0.4	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	0.3	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	0.3	2.4003	0.29	2.3064	0.30
31	Over-report by 100%	-0.5	2	-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	0.5	1.1801	0.46	1.0344	0.49
41	Over-report by 50%	-0.3333	1.5	*0	1	-0.2545	1.34
42	Under-report by 30%	0.4285	0.7	0.3491	0.74	0.3885	0.72
45	Under-report by 30%	0.4285	0.7	0.5508	0.64	0.4454	0.69
Others	Honest	0	1	-	-	-	-

* False Positive

TABLE VI

COMPARISON AMONG VARYING $a_{t_i,n}$, $\tilde{a}_{n(CVLR)}$, $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ AND $\bar{\tilde{a}}_{t_i,n(EADF)}$ OBTAINED FROM HARDWARE EXPERIMENTATION FOR THE SIZE OF 3 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 $\frac{1}{1+(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})}$	$\bar{\tilde{a}}_{t_i,n(EADF)}$	$\frac{1}{1+\bar{\tilde{a}}_{t_i,n(EADF)}}$
1	Over-report by 50%	All the time	-0.3333	1.5	-0.3368	-0.3335	1.51	-0.3371	1.51
2	Under-report by 60%	On-peak (From t_{16} to t_{39})	1.5	0.4	$0.0019851 \approx 0$	1.5120	0.40	1.5001	0.40
3	Honest	All the time	0	1	$0.0044069 \approx 0$	$0.0014501 \approx 0$	1	0	1

B. Varying Anomaly Coefficients

Besides, we also conduct 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 where $\nu \in (0, 0.95)$) 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. The varying under-reporting/over-reporting scenario for the hardware experimentation (size of 3 consumers) is set up as shown in Table VI.

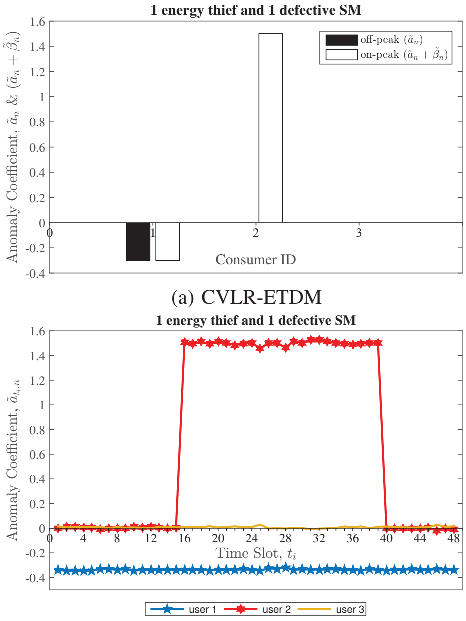
Here, values of $a_{t_i,n}$ are exact settings from the dataset, whereas, $\tilde{a}_{n(CVLR)}$ and $\tilde{\beta}_{n(CVLR)}$ are the estimated values obtained by the CVLR-ETDM. Values of $\bar{\tilde{a}}_{t_i,n(EADF)}$ are the average values of the computed $\tilde{a}_{t_i,n(EADF)}$ obtained by Enhanced 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). Figs. 3(a) and 3(b) depict the values of $\tilde{a}_{n(CVLR)}$ and $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ 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 VI.

As shown in Figs. 3(a) and 3(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 $\bar{\tilde{a}}_{t_i,1(EADF)} = -0.3371)$. On the other hand, consumer 2 under-reports what was consumed only during on-

TABLE VII
COMPARISON AMONG VARYING $a_{t_i,n}$, $\tilde{a}_{n(CVLR)}$, $(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})$ AND $\bar{\tilde{a}}_{t_i,n(EADF)}$ OBTAINED FROM [15] FOR THE 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 $\frac{1}{1+(\tilde{a}_{n(CVLR)} + \tilde{\beta}_{n(CVLR)})}$	$\bar{\tilde{a}}_{n(EADF)}$	$\frac{1}{1+\bar{\tilde{a}}_{n(EADF)}}$
1	Under-report by 60%	All the time	1.5	0.4	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	0.5	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	0.5	$0.0461 \approx 0$	1.0871	0.48	1.0000	0.5
19	Under-report by 70%	All the time	2.3333	0.3	2.4433	2.4689	0.29	2.3306	0.30
22	Under-report by 60%	On-peak (From t_{16} to t_{39})	1.5	0.4	$0.0461 \approx 0$	1.5869	0.39	1.4995	0.40
23	Under-report by 50%	On-peak (From t_{16} to t_{39})	1	0.5	$0.0415 \approx 0$	1.0547	0.49	0.9997	0.50
27	Under-report by 60%	On-peak (From t_{16} to t_{39})	1.5	0.4	$0.0500 \approx 0$	1.6263	0.38	1.5209	0.40
32	Under-report by 70%	On-peak (From t_{16} to t_{39})	2.33	0.3	$0.0411 \approx 0$	2.4679	0.29	2.3250	0.30
35	Under-report by 50%	On-peak (From t_{16} to t_{39})	1	0.5	$0.0378 \approx 0$	1.0662	0.48	0.9977	0.50
37	Over-report by 50%	All the time	-0.3333	1.5	-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	0.5	1.0946	1.0806	0.48	0.9974	0.50
Others	Honest	All the time	0	1	-	-	-	-	

* False Positive



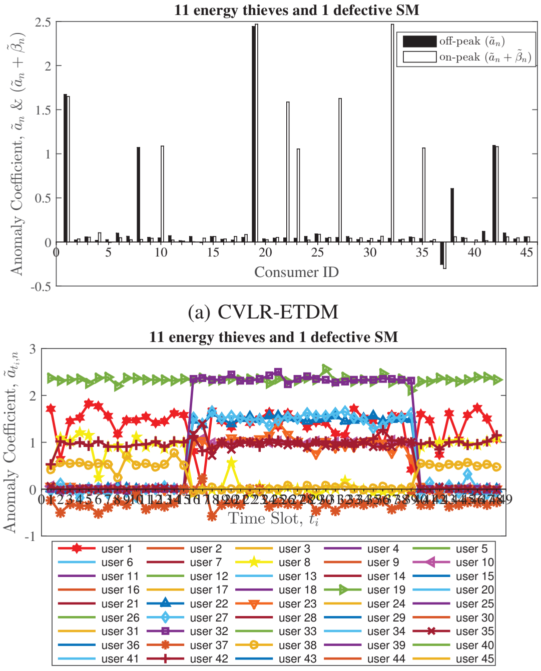
(a) CVLR-ETDM

(b) Enhanced ADF

Fig. 3. Values of anomaly coefficient obtained by CVLR-ETDM and Enhanced ADF from the test rig when $a_{t_i,n}$ is varying (size of 3 consumers)

peak period (i.e., $\tilde{a}_{2(CVLR)} = 0$, $(\tilde{a}_{2(CVLR)} + \tilde{\beta}_{2(CVLR)}) > 0.05$ while $(\bar{\tilde{a}}_{t_{16},2(EADF)}, \bar{\tilde{a}}_{t_{17},2(EADF)}, \cdots, \bar{\tilde{a}}_{t_{39},2(EADF)} > 0.05$, where $\bar{\tilde{a}}_{t_i,2(EADF)} = 1.5001)$). Consumer 3 is identified as honest as $\tilde{a}_{3(CVLR)} = 0$, $(\tilde{a}_{3(CVLR)} + \tilde{\beta}_{3(CVLR)}) = 0$ while $(\bar{\tilde{a}}_{t_1,3(EADF)}, \bar{\tilde{a}}_{t_2,3(EADF)}, \cdots, \bar{\tilde{a}}_{t_{48},3(EADF)} = 0)$. Similarly, both CVLR-ETDM and Enhanced ADF produce almost the same results when the amount of TLs is negligible.

To assess the scalability of the proposed anomaly detection frameworks under varying cheating/malfunctioning scenario in the presence of TLs and measurement noise, we consider a NAN of 45 consumers. Five-month half-hourly smart energy data (i.e., $D = 150$) from the Irish Smart Energy [15] 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 VII.



(a) CVLR-ETDM

(b) Enhanced ADF

Fig. 4. Values of anomaly coefficient obtained by CVLR-ETDM and Enhanced ADF from [15] when $a_{t_i,n}$ is varying (size of 45 consumers)

By referring to the results from Figs. 4(a) and 4(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 $(\bar{\tilde{a}}_{t_1,n(EADF)}, \bar{\tilde{a}}_{t_2,n(EADF)}, \cdots, \bar{\tilde{a}}_{t_{48},n(EADF)} > 0.05)$. 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 $(\bar{\tilde{a}}_{t_1,37(EADF)}, \bar{\tilde{a}}_{t_2,37(EADF)}, \cdots, \bar{\tilde{a}}_{t_{48},37(EADF)} < -0.05)$. Meanwhile, consumer 8 under-reports what was consumed only during off-peak period (i.e.,

TABLE VIII
PERFORMANCE COMPARISON BETWEEN LR-ETDM AND ADF

Scenario	No. of Consumer	Day	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 Positive

TABLE IX
PERFORMANCE COMPARISON BETWEEN CVLR-ETDM AND ENHANCED ADF

Scenario	No. of Consumer	Day	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 Positive

$\tilde{a}_{8(CVLR)} > 0.05$, $(\tilde{a}_{8(CVLR)} + \tilde{\beta}_{8(CVLR)}) = 0$, 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 Fig. 4(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 (12). Meanwhile, as presented in Table VII, 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 suggested 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 $\tilde{a}_{n(CVLR)} \approx 0$) as shown in Tables VI and VII based on the computed coefficients.

C. Performance Comparison Between LR-based and LP-based Anomaly Detection Frameworks

Table VIII 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 VIII that the detection rates (DR) of both LR-ETDM and ADF are 100% when the amount of TLs is negligible (i.e., case 1) or when TLs are non-existent (i.e., case 2). The results also suggested 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 positive (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 of TLs and noise on the detection analysis is not considered in the framework. On the contrary, the number of FP decreases in ADF when we observe consumers' energy consumption data over more days. In ADF, loss factor and error term capture the amount 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 IX 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 IX that Enhanced ADF requires more observation data (i.e., at least N days, where N is the number of consumers in the service area) as compared to CVLR-ETDM to achieve higher DR and lower FP. The table also suggested 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 amount of TLs and noise in the system, thereby improving the detection accuracy.

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 time, the number of FP increases. Therefore, it is important to take TLs and measurement noise/error into account in the design of anomaly detection framework to improve the robustness and accuracy of NTLs detection analysis. Besides identifying theft and irregularities in meter readings during specific off-peak/on-peak periods, our 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 NTLs detection during off-peak/on-peak periods only. However, the results are omitted here due to space constraints. The results are detailed in [11].

Table X shows a summary of all the proposed schemes. LR-ETDM and CVLR-ETDM which adopted MLR do not consider TLs in the NTLs detection analysis, hence the detection rate is lower as compared to the LP-based ADF and Enhanced ADF in the 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 [17]. In

TABLE X
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
LR-ETDM	MLR	✓	✗	✗	✗
CVLR-ETDM	MLR with Categorical Variables	✓	✓	✗	✗
ADF	LP	✓	✗	✗	✓
Enhanced ADF	LP	✓	✓	✓	✓

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 (3) and (7). The MLR-based anomaly detection schemes require less metered data as compared to LP-based ones to detect constant and varying NTLs. 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 types of NTLs.

V. CONCLUSION

In this work, we have constructed an advanced metering infrastructure test rig in the laboratory to perform comparison studies on our previously proposed linear regression-based and linear programming-based anomaly detection frameworks in smart grid environment. The detection frameworks are able to detect the fraudulent consumers who commit energy theft and pinpoint the localities of metering defects, with the goal to minimize revenue losses and costs incurred due to non-technical losses in smart grids. Any non-zero anomaly coefficient is indicative of meter irregularities. Results from both hardware experimentation and simulations show that linear regression-based anomaly detection framework is able to detect electricity pilfering and metering defects without requiring large volume of dataset. However, we found out that it might become unstable in the presence of technical losses or measurement noise in larger service area. Therefore, a new anomaly detection framework which is based on linear programming is designed to take the impact of technical losses and noise into consideration for more accurate non-technical losses detection. The proposed linear programming-based detection framework is still able to detect 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. In addition, it can estimate the amount of technical losses based on the knowledge of the distribution network and measurements at the data collector. The proposed frameworks differ from the existing work because regression and linear programming analysis are:

1) Not restricted by the dimension of the metered data.
2) Robust against non-malicious factors.
3) Protected against contamination attacks.
4) Capable of estimating the amount of NTL/TLs based on a small volume of energy consumption data samples regardless of the types of consumer.

These attributes lead to improved practicality and greater flexibility in residential/commercial energy fraud detection. Furthermore, the proposed frameworks can be extended easily to accommodate more consumers for anomaly detection.

REFERENCES

[1] L. Northeast Group, “$96 billion is lost every year to electricity theft,” 2017.

[2] P. Antmann, “Reducing technical and nontechnical losses in the power sector.” 2009.

[3] S. McLaughlin, D. Podkuiko, and P. McDaniel, “Energy theft in the advanced metering infrastructure,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6027 LNCS, pp. 176–187, 2010.

[4] J. Nagi, K. S. Yap, F. Nagi, S. K. Tiong, S. P. Koh, and S. K. Ahmed, “NTL detection of electricity theft and abnormalities for large power consumers in TNB malaysia,” in *2010 IEEE Student Conference on Research and Development (SCOReD)*, pp. 202–206, Dec 2010.

[5] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, “Decision tree and SVM-based data analytics for theft detection in smart grid,” *IEEE Trans. on Industrial Informatics*, vol. 12, no. 3, pp. 1005 – 1016, 2016.

[6] V. Badrinath Krishna, G. A. Weaver, and W. H. Sanders, “PCA-based method for detecting integrity attacks on advanced metering infrastructure,” in *Proceedings of the 12th International Conference on Quantitative Evaluation of Systems - Volume 9259*, QEST 2015, pp. 70–85, Springer-Verlag New York, Inc., 2015.

[7] V. B. Krishna, K. Lee, G. A. Weaver, R. K. Iyer, and W. H. Sanders, “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, June 2016.

[8] P. Jokar, N. Arianpoo, and V. C. M. Leung, “Electricity theft detection in AMI using customers’ consumption patterns,” *IEEE Trans. on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[9] S. Salinas, M. Li, and P. Li, “Privacy-preserving energy theft detection in smart grids,” in *Annual IEEE Communications Society Conf. on Sensor, Mesh and Ad Hoc Communications and Networks workshops*, vol. 1, pp. 605–613, 2012.

[10] S. C. Yip, K. S. Wong, W. P. Hew, M. T. Gan, R. C. Phan, and S. W. Tan, “Detection of energy theft and defective smart meters in smart grids using linear regression,” *Int. J. of Electrical Power and Energy Systems*, vol. 91, pp. 230–240, 2017.

[11] S. C. Yip, W. N. Tan, C. K. Tan, M. T. Gan, and K. S. Wong, “An anomaly detection framework for identifying energy theft and defective meters in smart grids,” *Int. J. of Electrical Power and Energy Systems*, vol. 101, no. March, pp. 189–203, 2018.

[12] *NJ-series machine automation controller database connection CPU unit*, 2017 (accessed October 21, 2017). [Online]. Available: https://www.valin.com/sites/default/files/asset/document/Omron-NJ-Database-CPU-Brochure.pdf.

[13] S. C. Yip, C. K. Tan, W. N. Tan, M. T. Gan, and A. H. A. Bakar, “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, June 2017.

[14] M. Yusoff, A. Busrah, M. Mohamad, and T. A. Mau, “A simplified approach in estimating technical losses in distribution network based on load profile and feeder characteristics,” in *PECon 2008 - 2008 IEEE 2nd International Power and Energy Conf.*, pp. 1661–1665, 2008.

[15] *Irish Social Science Data Archive (ISSDA)*, 2009 (accessed August 9, 2016). [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[16] S. A. Salinas and P. Li, “Privacy-preserving energy theft detection in microgrids: A state estimation approach,” *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 883–894, 2015.

[17] A. H. Studenmund, *Using Econometrics : A Practical Guide*. Pearson Education Limited, 6th ed., 2014.