Electrical Power and Energy Systems 91 (2017) 230–240

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Detection of energy theft and defective smart meters in smart grids using linear regression

Sook-Chin Yip [a,b,*], KokSheik Wong [c], Wooi-Ping Hew [a], Ming-Tao Gan [b], Raphael C.-W. Phan [b], Su-Wei Tan [b]

[a] UM Power Energy Dedicated Advanced Center (UMPEDAC), University of Malaya, Malaysia
[b] Faculty of Engineering, Multimedia University, Malaysia
[c] School of Information Technology, Monash University Malaysia, Malaysia

ARTICLE INFO

Article history:
Received 13 August 2016
Received in revised form 21 November 2016
Accepted 13 April 2017

Keywords:
Energy theft detection
Defective meter detection
Smart Grid
Linear regression
Categorical variable

ABSTRACT

The utility providers are estimated to lose billions of dollars annually due to energy theft. Although the implementation of smart grids offers technical and social advantages, the smart meters deployed in smart grids are susceptible to more attacks and network intrusions by energy thieves as compared to conventional mechanical meters. To mitigate non-technical losses due to electricity thefts and inaccurate smart meters readings, utility providers are leveraging on the energy consumption data collected from the advanced metering infrastructure implemented in smart grids to identify possible defective smart meters and abnormal consumers' consumption patterns. In this paper, we design two linear regression-based algorithms to study consumers' energy utilization behavior and evaluate their *anomaly coefficients* so as to combat energy theft caused by meter tampering and detect defective smart meters. Categorical variables and *detection coefficients* are also introduced in the model to identify the periods and locations of energy frauds as well as faulty smart meters. Simulations are conducted and the results show that the proposed algorithms can successfully detect all the fraudulent consumers and discover faulty smart meters in a neighborhood area network.

1. Introduction

Energy theft, which is also referred to as non-technical loss (NTL) has been a daunting problem for all utility providers (UPs) in the conventional power grid system. NTLs are generally related to energy theft and consumers fraudulent behavior in which there exist a number of methods to deliberately defraud the UPs [1]. NTLs may introduce a series of additional losses, such as reduction in grid reliability and damage to the grid infrastructure. NTLs include meter tampering, meter bypassing, meter switching, tapping on secondary voltages, error in computation of technical losses, defective meters, errors and delay in meter reading and billing, unpaid billing, etc. [2–4]. The latest estimates indicate that UPs suffer from losses up to six billion dollars annually due to energy fraud in the United State alone [5]. In recent years, Smart Grid (SG) is being globally introduced to replace its antiquated predecessor to address some of these issues. One significant feature of SG infrastructure is the replacement of the conventional mechanical meters by smart meters (SMs) in Advanced Metering Infrastructure (AMI).

The introduction of SGs and SMs may contribute to a significant cutback in NTLs by minimizing some types of losses [2,6]. However, the SG, AMI in particular, raises new security risks [5,7–11]. Specifically, AMI can be exploited by the adversaries to perform a number of attacks for manipulating the energy utilization statistics because SMs are vulnerable to more types of attack such as network-borne attacks. In addition, consumers' consumption data may be compromised at three different stages, namely during transmission to UP, while it is being recorded, or after it is stored [12]. Since the conventional methods for mitigating NTLs impose high operational costs (e.g., on-site inspection where extensive deployment of human resources is involved [13,14]), this paper

Abbreviations: NTL, non-technical loss; UP, utility provider; SG, Smart Grid; SM, smart meter; AMI, Advanced Metering Infrastructure; NAN, neighborhood area network; MLR, multiple linear regression; RFID, radio frequency identification; SVM, support vector machine; GA, genetic algorithm; LUD, LU decomposition; DS, distribution station; TL, technical loss; LSE, linear system of equations; TOU, Time-of-Use.

* Corresponding author at: UM Power Energy Dedicated Advanced Centre (UMPEDAC), Level 4, Wisma R&D University of Malaya, Jalan Pantai Baharu 59990 Kuala Lumpur, Malaysia.

E-mail addresses: scyip@mmu.edu.my (S.-C. Yip), wong.koksheik@monash.edu (K. Wong), wphew@um.edu.my (W.-P. Hew), mtgan@mmu.edu.my (M.-T. Gan), raphael@mmu.edu.my (R.C.-W. Phan), swtan@mmu.edu.my (S.-W. Tan).

http://dx.doi.org/10.1016/j.ijepes.2017.04.005
0142-0615/© 2017 Elsevier Ltd. All rights reserved.

aims to reduce the operational costs of UPs by detecting NTL activities.

In this paper, we propose two linear regression-based algorithms to identify the locations of defective SMs and malicious SMs which are compromised by energy thieves to falsify readings (i.e., data attacks [15]) in the neighborhood area network (NAN). The key idea is to adopt multiple linear regression (MLR) for estimating and evaluating consumers' *anomaly coefficients* based on the reported consumers' energy consumption data. MLR is chosen because it adopts characteristic analysis, which attempts to model the consumers' energy consumption behavior for consideration [16]. Therefore, any anomalies not following the utilization trend may be indicative of energy thefts or metering defects. MLR analysis is especially attractive as it is able to accurately reveal not only the locations of energy thieves and defective SMs, but also the amount of energy theft/loss.

2. Related work

Broadly, energy theft detection techniques, including those that are widely implemented in both conventional power grids and SGs, may be grouped into two categories, namely *state-based detection* and *classification-based detection*.

2.1. State-based detection

This method utilizes monitoring state through mutual inspection [12], wireless sensor networks [15], control units [17], radio frequency identification (RFID) [18] and distribution transformers [19] to identify fraud in power system.

As detailed in [12], Xiao et al. proposed three inspection algorithms to identify malicious SMs in a neighborhood. First, they developed a basic scanning method. Then, they designed a binary tree-based method for inspection when the *malicious SMs* to *honest users* ratio is high, and finally employed an adaptive tree-based method to leverage on the advantages of both the scanning and binary tree algorithms. However, adding an extra meter for each consumer/provider will significantly increase the cost. Meanwhile, the authors in [15] designed an AMI Intrusion Detection System (AMIDS), which utilizes information fusion to combine the consumption and sensors data from a SM to model and identify fraud-related behavior more accurately. In [17], consumers consumption data is compared with the feeder input level. Both individual and aggregated consumption are also compared against the feeder details to detect consumption anomalies. However, their proposal can only detect a small region of electricity theft but not the exact location of fraud. Khoo and Cheng [18] proposed a system that incorporated RFID technology to assist the UPs in ammeter inventory management and mitigate energy theft. Although RFID technology can be implemented to identify electricity theft, UPs have to pay extra cost to install the system. In [19], the author 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 distribution transformers 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.

2.2. Classification-based detection

The key idea of this approach is to identify consumers' energy consumption anomalies based on testing datasets consisting of the normal and attack class samples using machine learning [20]. Han et al. [2] designed a NTL fraud detection scheme by using the approximated difference between the actual consumed electricity and billing electricity. On the other hand, Nizar et al. designed a feature selection-based approach to extract features from consumers' behaviors for further analysis [21] to find optimal subsets of features in establishing the load profiles, which describe consumers' energy consumption patterns over a period of time. An attacker model for anomaly detector in meter data management is developed by Mashima and C'ardenas to detect energy theft [22]. In [23], Nagi et al. studied consumers' behaviors and proposed a Remote Meter Abnormality Detection System to detect illegal and abnormal energy consumption trends using meter event logs and remote meter reading. In a different work [24], they proposed a fraud detection framework using Support Vector Machine (SVM). Their proposal chose some suspicious consumers in advance for on-site inspection for fraud based on the abnormal power consumption behavior. SVM is trained to extract features and generate fraud detection model. They also designed a hybrid method for NTL analysis by incorporating Genetic Algorithm (GA) and SVM [25]. Similar to [24], the algorithm selected suspicious consumers for inspection. Then, GA provides an increased convergence and optimized SVM hyper-parameters. Meanwhile, Depuru et al. [26] introduced high performance computing to speed up the energy theft detection through data encoding without compromising the quality of data. The encoded data are then classified to discover the electricity pilfering using SVM and Rule Engine-based algorithms. The authors in [27] shortlisted area with high probability of theft using distribution transformers. Then, their proposal identified the suspicious consumers by observing irregularities of consumption patterns using SVM. The SVM-based energy theft detection schemes [24–27] usually require a large volume of training data with load profiles collected from SMs to extract features from historical data.

Besides, it is crucial to preserve consumers' privacy while detecting energy theft in SGs as detailed in [28,29]. In their paper, Salinas et al. proposed a LU decomposition-based (LUD) algorithm to solve a linear system of equations for consumers' honesty coefficients while ensuring consumers' privacy. However, their proposal 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, the authors have to change the time granularity. Nevertheless, it might not be practical to *reduce the sampling period or time granularity* indefinitely due to the memory size of SM.

To address some of the limitations of previous work, linear regression-based schemes for identifying energy thefts and defective SMs which are not restricted by the dimension of consumers' power consumption data as well as its time granularity are proposed in this work.

3. Architecture of smart grid in neighborhood area network

Here, we present the electrical and communication network architectures considered in this paper. In AMI, the electrical and communication networks overlay each other and all electrical and communication flows are bidirectional [30]. According to the surveys of SG [9,11], the architecture of SG in a neighborhood area network (NAN) can be illustrated in Fig. 1. Further details on *Electrical network* and *Communication network* will be provided below.

3.1. Electrical network

Similar to the conventional electrical grid system, the power supply of SG in a NAN is usually serviced by the same UP. The UP builds a distribution station (DS), which is also known as fuse

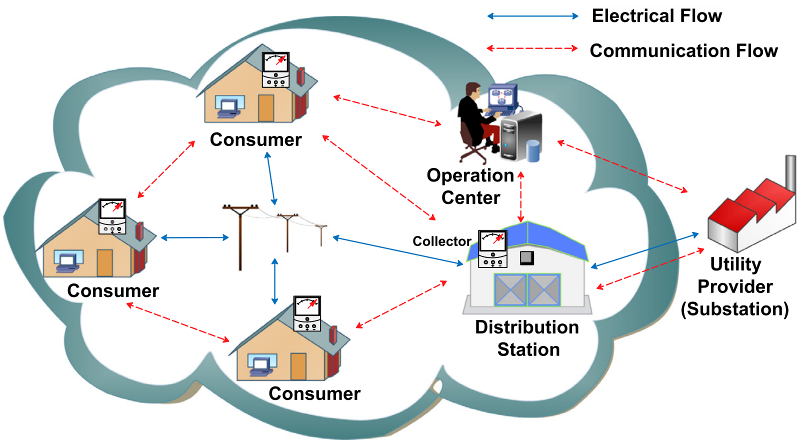


Fig. 1. The architecture of smart grid in neighborhood area network.

box [31] within every neighborhood. The DS acts like an 'electricity router' to distribute power from the substation to all the consumers in the neighborhood. A master SM, known as the **collector** is endowed inside the DS to measure the aggregated power supply from the UP to all consumers in the NAN 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 $n \in \{1, 2, \ldots, N\}$, UP installs a SM at each consumer's household. The n-th SM 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 $t_i - t_{i-1}$. Thus, we have [2,7]

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

where c_{t_i} denotes the total energy supplied by the UP to all consumers (i.e., N of them) in the NAN, λ denotes the technical losses (TLs), as well as the reduced meter readings due to energy thefts (i.e., θ) and faulty SMs (i.e., γ).

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 t_i, denoted by y_{t_i}, is computed as:

$$y_{t_i} = c_{t_i} - \sum_{n=1}^{N} p_{t_{i,n}} = \lambda + \theta + \gamma. \tag{2}$$

3.2. Communication network

The SMs installed in households, collector, operation center and DS form a *neighborhood area network* (NAN). In a NAN, UP relies on an operation center to monitor the DS and distribution networks. The communications among the SMs and the collector are conducted in a wireless manner while the communications among the collector, operation center, DS and substation are conducted via wired medium such as power feeder line [31]. In our model, we assume all consumers premises are endowed with a SM. Therefore, we do not consider the effect caused by consumers without a SM.

4. Linear regression model for detecting energy theft and defective smart meters

We present the mathematical model for detecting energy theft and defective SMs in a NAN. Suppose that UP equips a SM at each household to record the electricity consumption at some predefined time intervals. Besides, a collector is installed inside the DS such that the collector can measure the aggregated power supply from the UP to the service area.

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 collector, respectively, at time interval $t_i \in T$. We further define an *anomaly coefficient*, denoted by a_n, for each consumer such that $a_n = 0$ if consumer n is honest in reporting his/her energy consumption. Therefore, $(a_n + 1)p_{t_{i,n}}$ gives the cumulative energy consumption reported by consumer n at t_i. Since 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 [28], the following can be formulated:

$$(a_1 + 1)p_{t_{i,1}} + (a_2 + 1)p_{t_{i,2}} + \cdots + (a_n + 1)p_{t_{i,n}} = c_{t_i}. \tag{3}$$

To facilitate the discussion, Eq. (3) is re-arranged as:

$$a_1 p_{t_{i,1}} + a_2 p_{t_{i,2}} + \cdots + a_n p_{t_{i,n}} = c_{t_i} - \sum_{n=1}^{N} p_{t_{i,n}}. \tag{4}$$

Similar to Eq. (2), the right hand side of Eq. (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 our model does not consider TLs (in which its percentage is denoted by λ) in the SGs. TLs occur during power distribution and transmission, which involve substations, transformers and line-related losses [21]. TLs also occur due to dynamic environment factors (e.g., temperature) and are caused by the low voltage power lines as well as intrinsic inefficiencies in the transformers [28]. Nonetheless, Sahoo et al. [32] 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, our proposed model can be adjusted accordingly by subtracting TLs from vector **y** as expressed in Eq. (2).

Our goal is to find all a_n in the linear system of equations (LSE) from Eq. (4) so as to evaluate the anomalous behavior of each consumer or reliability of SM endowed in each household. In particular, there are three possibilities:

$a_n = 0$**:** Consumer n is honest and does not cheat.
$a_n > 0$**:** Consumer n reports less energy consumption than what was consumed (i.e., energy theft).
$a_n < 0$**:** The n-th SM reports more than what was consumed (i.e., faulty SM).

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

$$\begin{cases} a_1 p_{t_{1,1}} + a_2 p_{t_{1,2}} + \cdots + a_N p_{t_{1,N}} = y_{t_1} \\ \vdots \\ a_1 p_{t_{T,1}} + a_2 p_{t_{T,2}} + \cdots + a_N p_{t_{T,N}} = y_{t_T} \end{cases} \tag{5}$$

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

$$\mathbf{Pa} = \mathbf{y} \tag{6}$$

where

$$\mathbf{P} = \begin{bmatrix} p_{t_{1,1}} & p_{t_{1,2}} & \cdots & p_{t_{1,N}} \\ p_{t_{2,1}} & p_{t_{2,2}} & \cdots & p_{t_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ p_{t_{T,1}} & p_{t_{T,2}} & \cdots & p_{t_{T,N}} \end{bmatrix},$$

$$\mathbf{a} = [a_1,\ a_2, \ldots,\ a_N]' \text{ and } \mathbf{y} = [y_{t_1},\ y_{t_2}, \ldots,\ y_{t_T}]'. \tag{7}$$

Here, the t_i-th row of $\mathbf{P}$ represents the data recorded by all N consumers at the t_i-th time interval. On the other hand, the n-th column of $\mathbf{P}$ denotes the data measured by the SM for consumer n over all t_i. In this model, $\mathbf{a}$ is a column vector consisting of anomaly coefficients $a_1,\ a_2, \ldots,\ a_N$.

We explain the scenario using a simple 2-consumer topology, namely consumer A and consumer B. As mentioned previously, if there are no energy thefts or defective SMs at $t_i,\ y_{t_i} = 0$ in Eq. (2) and then Eq. (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 $a_B = 0$ imply that consumer A cheats on the SM readings while consumer B is honest.

5. Estimating anomaly coefficients using linear regression

In the following subsections, we develop two algorithms to solve the LSE for the anomaly coefficients in Eq. (6) using linear regression. Our objective is to enable the collector to reveal the locations of energy thieves and/or faulty SMs.

5.1. Multiple linear regression

We first develop 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**, to detect energy thieves and defective SMs. Linear regression is a modeling technique utilized to explicitly describe the relationship between a continuous-valued response Y_i and linear predictors $p_{t_{i,1}},\ p_{t_{i,2}}, \ldots,\ p_{t_{i,N}}$. 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 be estimated using a range of independent variables [33,34]. Here, y_{t_i} as defined in Eq. (2) is viewed as the realization of a normally distributed random variable $Y_i \sim N(d_{t_i}, \sigma^2)$, where

$$d_{t_i} = \alpha + \sum_{n=1}^{N} a_n p_{t_{i,n}}. \tag{8}$$

Eq. (8) defines a hyper-plane [35], where the parameter α (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 our work, we set $\alpha = 0$ due to the assumption that the response is entirely dependent on the predictors.

An important characteristic of the linear regression-based model (i.e., Eq. (8)) is that it is additive [35]. 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 our model as expressed in Eq. (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, we assume that consumers steal energy or SMs are damaged all the time. This assumption may be unfeasible, and therefore later in Section 6 we will introduce an enhanced model which captures the changes of the estimated anomaly coefficients to identify the period of the fraud and/or metering defects.

It has been shown in [35] that the maximum likelihood estimate of the coefficients $\mathbf{a}$ are those that minimize the residual sum of squares between y_{t_i} and d_{t_i}. If $\mathbf{P}$ is of full column rank, then $\mathbf{a}$ is given by:

$$\mathbf{a} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{y}. \tag{9}$$

5.2. Student's t-statistic and two-tailed p-value approach

As mentioned in the previous section, Eq. (9) is introduced to compute the absolute value of all anomaly coefficients, $\mathbf{a}$. However, there is no objective way to determine whether the value of the computed anomaly coefficient is 0 or 1. In linear regression, 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

$$\begin{cases} H_0 : a_n = 0 \\ H_1 : a_n \neq 0 \end{cases}, \tag{10}$$

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 [35,36]. *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 model. Meanwhile, *p*-value is a function of the *t*-statistic that is utilized for comparing the probability of rejecting H_o 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 [37], but in our work, we choose the latter 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.

5.3. The LR-ETDM algorithm

In this section, we detail the LR-ETDM algorithm. Here, we assume a constant scenario where the fraudulent consumers always steal energy and the defective SMs always report more than what the corresponding consumers actually consumed.

The flowchart as shown in Fig. 2 summarizes the LR-ETDM scheme. Assume that the collector labels the SM of all consumers in the service area of interest from 1 to *N*. SM_n 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 algorithm commences by computing the discrepancies between the total power supplied by the UP (i.e., 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 $t_i \in T$. Then, a LSE consisting of consumers' reported load data, anomaly coefficients and the differences in reading is formed as expressed by Eq. (5). In this work, we use the `fitlm` function built in the Statistics Toolbox of Matlab R2014b 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., α in Eq. (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., $\forall n \in N$) are found using linear regression method. Based on the *p*-values and estimated a_n, we can pinpoint the locations of energy frauds and faulty SMs.

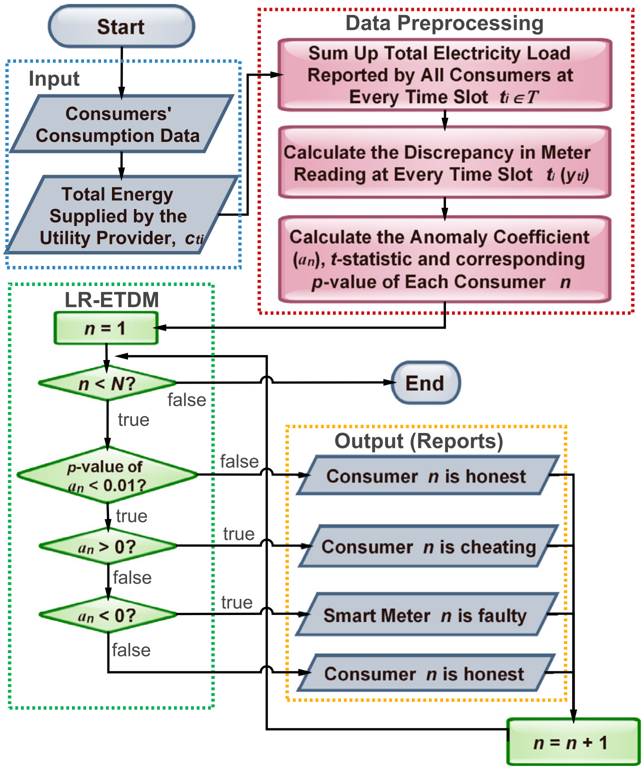

Start
Input
Consumers' Consumption Data
Total Energy Supplied by the Utility Provider, c_{ti}
Data Preprocessing
Sum Up Total Electricity Load Reported by All Consumers at Every Time Slot $t_i \in T$
Calculate the Discrepancy in Meter Reading at Every Time Slot t_i (y_{ti})
Calculate the Anomaly Coefficient (a_n), t-statistic and corresponding p-value of Each Consumer n
LR-ETDM
n = 1
n < N?
false
true
End
p-value of a_n < 0.01?
false
true
a_n > 0?
true
false
a_n < 0?
true
false
Output (Reports)
Consumer n is honest
Consumer n is cheating
Smart Meter n is faulty
Consumer n is honest
n = n + 1

Fig. 2. Flow chart of LR-ETDM.

For every consumer $n \in N$, 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 a 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 thieves when 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 earlier, linear regression explicitly assumes that the anomaly coefficients a_n do not change throughout the period of observation [38]. 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 6, we design an enhanced algorithm to reveal the locations and periods (i.e., during peak, off-peak of a day or whole day) of energy theft or device failure by introducing *categorical variables* in linear regression.

6. Estimating variable anomaly coefficients using categorical variable method

In LR-ETDM, we assumed that the anomaly coefficients, $a_1, a_2, \ldots, a_N$ are constant. However, it is possible that the rate at which the fraudulent consumers steal electricity is variable when they commit energy theft [28]. 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 [8] tariffs. Therefore, consumers will be motivated to reduce energy costs by shifting some energy-intensive loads to off-peak hours or tampering 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. Hence, we propose another algorithm, **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**), by introducing *categorical variables* in linear regression through *dummy coding* to resolve the varying cheating problem.

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) [39]. 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 [40,41].

In our work, we include the categorical variables, x_i for $i = 1,\ 2, \ldots,\ N$ to categorize the time of fraud or metering defect of consumers $1,\ 2, \ldots,\ N$. 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 and each of them commits energy theft independently. 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*, $\boldsymbol{\beta}$ to the regression equation as follows:

$$\begin{gathered} a_1 p_{t_{1,1}} + \cdots + a_N p_{t_{1,N}} + \beta_1 p_{t_{1,1}} x_1 + \cdots + \beta_N p_{t_{1,N}} x_N = y_{t_1} \\ \vdots \\ a_1 p_{t_{T,1}} + \cdots + a_N p_{t_{T,N}} + \beta_1 p_{t_{T,1}} x_1 + \cdots + \beta_N p_{t_{T,N}} x_N = y_{t_T}, \end{gathered}$$

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

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

$$a_1 p_{t_{o,1}} + \cdots + a_N p_{t_{o,N}} + \beta_1 p_{t_{o,1}} \cdot 0 + \cdots + \beta_N p_{t_{o,N}} \cdot 0 = y_{t_o},$$

whereby $p_{t_{o,n}}$ denotes the energy consumption reported by consumer n during off-peak hours at time interval $t_o \in \{08:00\ \text{P.M.},\ 08:30\ \text{P.M.}, \ldots,\ 07:30\ \text{A.M.}\}$. Note that the time granularity is 30 min. Thus, we have

$$a_1 p_{t_{o,1}} + \cdots + a_N p_{t_{o,N}} = y_{t_o}, \tag{11}$$

for $\forall\ t_o$.

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

$$\mathbf{P}^{\mathbf{off}} \mathbf{a} = \mathbf{y}^{\mathbf{off}}, \tag{12}$$

which is similar to Eq. (6). In Eq. (12), $\mathbf{a}$ 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, we can form another LSE to detect consumers who perpetrate theft during on-peak hours or faulty SMs as follows:

$$a_1 p_{t_{p,1}} + \cdots + a_N p_{t_{p,N}} + \beta_1 p_{t_{p,1}} \cdot 1 + \cdots + \beta_N p_{t_{p,N}} \cdot 1 = y_{t_p},$$

which can also be re-arranged as:

$$(a_1 + \beta_1) p_{t_{p,1}} + \cdots + (a_N + \beta_N) p_{t_{p,N}} = y_{t_p}, \tag{13}$$

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

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

$$\mathbf{P}^{\mathbf{peak}} (\mathbf{a} + \boldsymbol{\beta}) = \mathbf{y}^{\mathbf{peak}}, \tag{14}$$

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

The relationship between Eqs. (12) and (14) can be represented by partitioned matrices as follows:

$$\left[\begin{array}{c|c} \mathbf{P}^{\text{off}} & 0 \\ \hline \mathbf{P}^{\text{peak}} & \mathbf{P}^{\text{peak}} \end{array}\right] \left[\begin{array}{c} \mathbf{a} \\ \hline \boldsymbol{\beta} \end{array}\right] = \left[\begin{array}{c} \mathbf{y}^{\text{off}} \\ \hline \mathbf{y}^{\text{peak}} \end{array}\right]. \tag{15}$$

By applying Eq. (9), the maximum likelihood estimator of our regression coefficients are thus computed by:

$$\left[\begin{array}{c} \mathbf{a} \\ \hline \boldsymbol{\beta} \end{array}\right] = ((\mathbf{P}^{\text{aug}})'\mathbf{P}^{\text{aug}})^{-1}(\mathbf{P}^{\text{aug}})'\mathbf{y} \tag{16}$$

where

$$\mathbf{P}^{\text{aug}} = \left[\begin{array}{c|c} \mathbf{P}^{\text{off}} & 0 \\ \hline \mathbf{P}^{\text{peak}} & \mathbf{P}^{\text{peak}} \end{array}\right]. \tag{17}$$

By investigating the estimated $\mathbf{a}$ and $\boldsymbol{\beta}$, we can deduce whether the dishonest consumers are committing theft either all the time or only during a particular period in a day. The following seven scenarios describe the operation of Eqs. (12) and (14) to identify cheating consumers or faulty SMs that occur constantly or occasionally through dummy coding. The possible scenarios of each consumer (i.e., $n = 1,\ \ldots,\ N$) are summarized in Table 1.

- **Scenario 1:** Obviously, both a and β equal to 0 imply that each consumer is honest in his/her energy reporting.
- **Scenario 2:** When a is positive while $\beta = 0$, the sum of a and β is also positive. $\beta = 0$ indicates that the anomaly coefficient is constant throughout the observed period. Therefore, we can conclude that the consumer is cheating on his/her energy consumption during both off-peak and on-peak hours (all the time).
- **Scenario 3:** If a is negative and $\beta = 0$, the total of a and β 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 β is positive. The positive sum of a and β 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 β shows that there is a status change from non-cheating during off-peak to cheating during on-peak.

Table 1
Description of a, β and ($a + \beta$).

Scenario	a	β	$a + \beta$	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

- **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 β 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:** β 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.

6.2. The CVLR-ETDM algorithm

The flow chart in Fig. 3 shows the operations in 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 time attributes (i.e., $k = 2$), namely off-peak and on-peak. Therefore, one dummy variable (i.e., $k - 1 = 1$) is created for each consumer. In total, we have $2N$ 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 β_n is verified to test the significance of the coefficient given the other coefficients. If the p-value of β_n is less than 0.01, it means that the t-statistic is significant at a 1% level given the other coefficients. In other words, β_n is non-zero (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 Eq. (14) 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 β_n greater than 0.01 implies that a_n of consumer n is constant. That is, consumer n cheats or n-th SM is malfunctioning 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.

7. Performance evaluation

We conduct two series of simulations in Matlab R2014b to evaluate the performance of our proposed LR-ETDM and CVLR-ETDM schemes. Specifically, two scenarios are considered, namely, fraudulent consumers steal at a *fixed rate* (constant anomaly coefficient) and *variable rate* (variable anomaly coefficient).

According to Jokar et al. [27] and Sahoo et al. [32], 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 Energy Trial denoted by $\mathbf{P}$, are extracted from [42] in our study. The SM electricity trial dataset was released by Electric Ireland and Sustainable Energy Authority of Ireland (SEAI) in March 2012. It consists of half-hourly energy usage reports for over 5000 Irish residential and commercial premises during 2009 and 2010. Consumers who took part in the trial had a SM endowed in their premises. Since the participation in this trial is voluntary, it is justifiable to assume that all samples are collected from honest consumers who reported the actual utilization. In addition, based on the trial dataset in [42], three types of malicious samples for each half-hourly sample $\mathbf{P_n} = \{p_{t_{1,n}}, p_{t_{2,n}}, \ldots, p_{t_{48,n}}\}$, for time $t_i = t_1, t_2, \ldots, t_{48}$ are generated:

1. $h_1(p_{t_{i,n}}) = vp_{t_{i,n}}, \; v = ([0, 0.9] \cup [1.1, 2.0])$;
2. $h_2(p_{t_{i,n}}) = \delta_{t_i} p_{t_{i,n}}$
   $$\delta_{t_i} = \begin{cases} v, & start < t_i < end \\ 1, & \text{otherwise} \end{cases}$$
   where v is as defined in (1) above, *start* and *end* are the starting and ending time of either *on-peak* or *off-peak* period;
3. $h_3(p_{t_{i,n}}) = \eta_{t_i} p_{t_{i,n}}$
   $$\eta_{t_i} = \begin{cases} 0, & start < t_i < end \\ 1, & \text{otherwise} \end{cases}$$
   where *start* and *end* are the starting and ending time of either *on-peak* or *off-peak* period.

In the first scenario, h_1 multiplies the meter readings by the same randomly chosen percentage, which remains constant (i.e., fixed rate). When fraudulent consumer steals energy at a fixed rate, he/she consistently reports a fraction of his/her consumed energy (e.g., 50% of the actual consumed data). In h_2, the energy thief cheats only during a certain period in a day (i.e., either on-peak or off-peak only). For instance, the fraudulent consumer reports 40% less than the consumed data during on-peak hours and reports the actual consumption data during off-peak hours. Using h_3, the SM sends zero reading or does not have measurements during a certain period in a day. In the simulations, we assume that defective SMs always report more than what the consumers actually consumed (i.e., $v = [1.1, 2.0]$).

As discussed previously in Section 4, our model does not consider technical losses (TLs) in the SGs. Nevertheless, TLs can be computed by observing the data from distribution transformers and the current readings collected by smart or conventional power meters [32]. Therefore, once the TLs are calculated, the proposed model can be adjusted accordingly by subtracting TLs from vector $\mathbf{y}$ as expressed in Eq. (2).

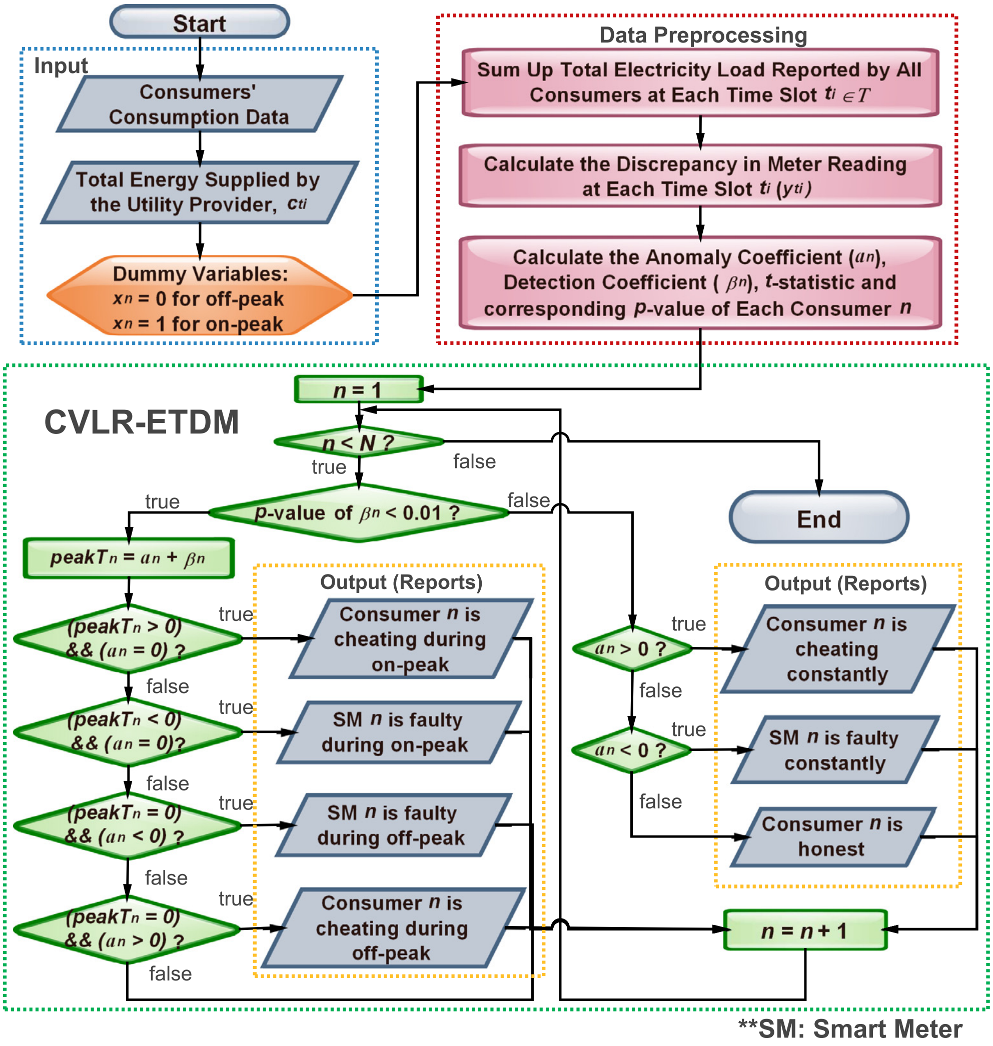

Start
Input
Consumers' Consumption Data
Total Energy Supplied by the Utility Provider, c_{t_i}
Dummy Variables: x_n = 0 for off-peak x_n = 1 for on-peak
Data Preprocessing
Sum Up Total Electricity Load Reported by All Consumers at Each Time Slot $t_i \in T$
Calculate the Discrepancy in Meter Reading at Each Time Slot t_i (y_{t_i})
Calculate the Anomaly Coefficient (a_n), Detection Coefficient (β_n), t-statistic and corresponding p-value of Each Consumer n
CVLR-ETDM
n = 1
n < N ?
true
false
p-value of β_n < 0.01 ?
true
false
End
peakT_n = a_n + β_n
Output (Reports)
(peakT_n > 0) && (a_n = 0) ?
Consumer n is cheating during on-peak
(peakT_n < 0) && (a_n = 0)?
SM n is faulty during on-peak
(peakT_n = 0) && (a_n < 0) ?
SM n is faulty during off-peak
(peakT_n = 0) && (a_n > 0) ?
Consumer n is cheating during off-peak
a_n > 0 ?
Consumer n is cheating constantly
a_n < 0 ?
SM n is faulty constantly
Consumer n is honest
n = n + 1
**SM: Smart Meter

Fig. 3. Flow chart of CVLR-ETDM.

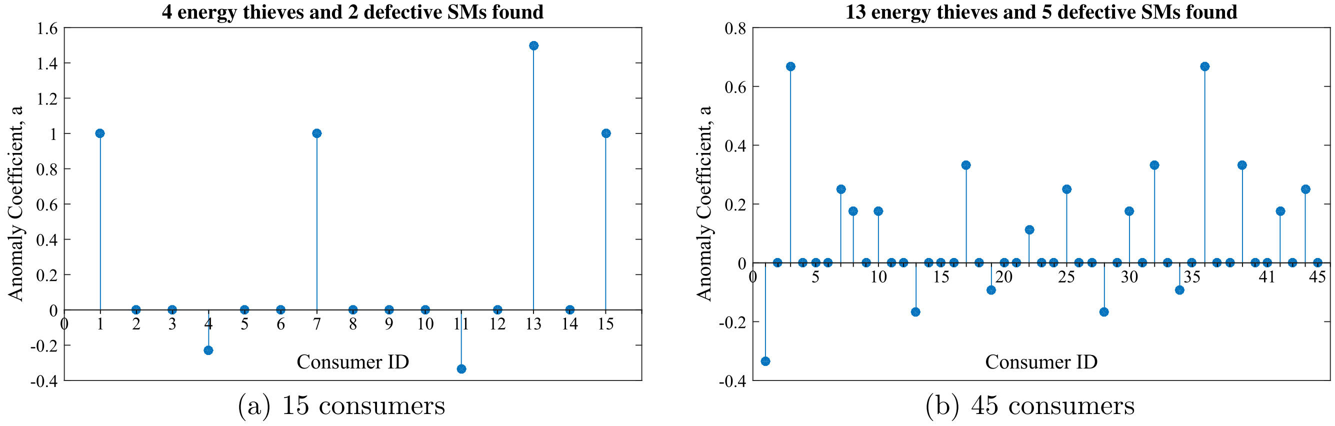
7.1. Constant anomaly coefficients

Here, we assume that the fraudulent consumers steal energy all the time and never stop cheating (i.e., h_1, where v = [0,0.9]). At the same time, some of the SMs are malfunctioning continuously (i.e., h_1, where v = [1.1,2.0]). Therefore, the rates of cheating as well as reporting more (due to malfunctioning) do not change and hence the anomaly coefficients are constant.

Service area of sizes 15 and 45 energy consumers are considered. Without loss of generality, we assume that 40% of the consumers are stealing energy and/or SMs are reporting more on their energy usage (i.e., SMs are out of order) constantly in the service area, and the time granularity is 30 min. Each energy thief n has an a_n in $[-0.5, 9] \setminus \{0\}$, depending on how much more they have reported or how much less they paid for the bill [28].

As shown in Fig. 4, the proposed LR-ETDM method can perform well for each of the cases we consider, i.e., when there are 15 and 45 consumers in the service area. In particular, in the case of 15 consumers, it is observed that there are six consumers who have anomaly coefficients which are not equal to 0 in Fig. 4(a). As shown in the figure, there are four energy thieves (i.e., consumer 1, 7, 13 and 15) who only report fraction of their energy consumption (i.e., $a_n > 0$). 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., $a_n < 0$). Based on these results, the collector can effectively detect all the energy thieves as well as the defective SMs, then computes how much less or more they have paid in their monthly bills. Besides, we can also easily identify the nine honest consumers in the service area who have $a_n = 0$. Similar result is observed in Fig. 4(b) for the case of 45 consumers. By isolating the consumers who have anomaly coefficients not equal to 0, we can effectively recognize the positions of energy thieves and defective SMs in the NAN.

Besides, we also conduct simulation by using LR-ETDM when some fraudulent consumers are cheating inconsistently and some of them are stealing energy constantly. Specifically, some of the dishonest consumers are stealing energy all the time and some of them are cheating on their energy consumption only during a certain period in a day. The results are presented in Fig. 5. As discussed earlier, the LR-ETDM algorithm becomes unstable under this scenario. It finds five cheating consumers and a faulty SM only but, by construction, there are five energy thieves and two faulty SMs. In fact, the scenario is setup as follows: consumer 1, 8, 13 and 15 are cheating constantly, the 4-th SM is out of order all the time, consumer 7 is cheating during on-peak while the 11-th SM is out of order only during peak-hours. However, LR-ETDM accuses the honest consumer 10 and 14 wrongly. Meanwhile, consumer 4, 7 and 13 are left unidentified.



(a) 15 consumers (b) 45 consumers

Fig. 4. Value of *a* obtained by LR-ETDM when *a* is constant.

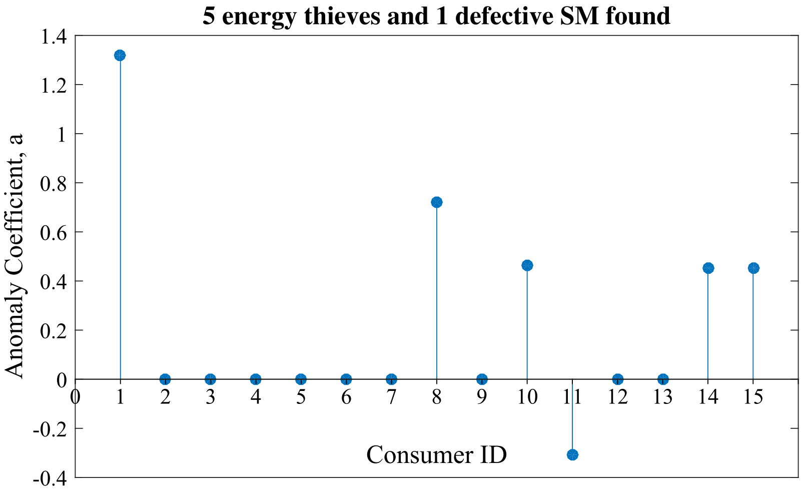


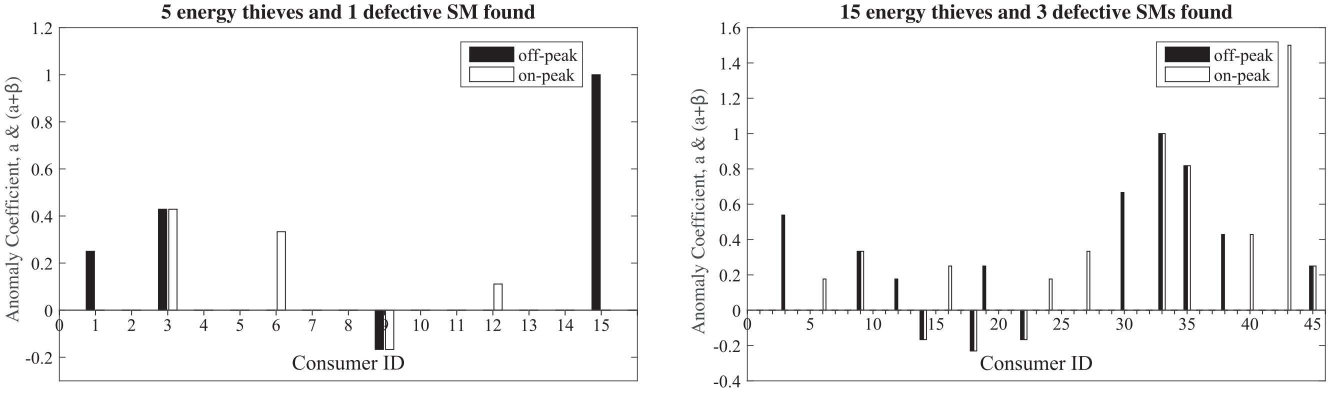
Fig. 5. Value of *a* obtained by LR-ETDM when *a* is variable.

7.2. Variable anomaly coefficients

Here, we conduct simulations for the situation when energy thieves cheat on their energy reporting and/or SMs are malfunctioning all the time/during a certain time (i.e., h_1, h_2 and h_3). The goal is to verify the viability of the proposed CVLR-ETDM in handling the consistent/inconsistent cheating and malfunctioning problems. We assume that each energy thief chooses a new anomaly coefficient uniformly and/or occasionally in $[-0.5, 9] \setminus \{0\}$ each time and 40% of the consumers and/or SMs in the NAN have a non-zero anomaly coefficient. Also, each fraudulent consumer commits energy theft during on-peak, off-peak or all the time. In the simulations, we observe consumers' power consumption data over two days to increase the number of observations so as to mitigate the effect of over-fitting [43].

Consider the results for 15 consumers. In Fig. 6, black bar represents off-peak period, $\mathbf{a}$ (i.e., variable anomaly coefficient) and white bar represents on-peak period, $\mathbf{a} + \boldsymbol{\beta}$ (i.e., variable anomaly coefficient). If white bar and black bar co-exist (i.e., constant anomaly coefficient), it implies that the energy frauds occur or defective meters exist all the time. Results in Fig. 6(a) suggest that there are five dishonest consumers and a faulty SM in the service area. In particular, consumer 1 and consumer 15 steal (i.e., $a > 0$ and $a + \beta = 0$, where $\beta = -a$) only during off-peak period (i.e., black bar) while consumer 6 and consumer 12 steal (i.e., $a = 0$ and $a + \beta > 0$) only during on-peak period (i.e., white bar). Meanwhile, consumer 3 is stealing all the time (i.e., black and white bar, $a > 0$, $a + \beta > 0$) during both off-peak and on-peak period. The 9-th SM is out of order all the time (i.e., black and white bar, $a < 0$, $a + \beta < 0$). In other words, if the consumer is stealing or the SM is defective all the time, the rates of cheating/malfunctioning do not change and hence the anomaly coefficients are constant. On the other hand, when consumer cheats inconsistently, the rates of cheating will change and hence the anomaly coefficients are variable. Based on these findings, the 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. Similar result is obtained for the case of 45 consumers, where the result is shown in Fig. 6(b).

Meanwhile, when a fraudulent consumer n attempts to send zero readings all the time or during a certain period in the day (i.e., scenarios h_1 when $v = 0$ or h_3), the p-value of the a_n is not a number (NaN). In such a case, the SM of the dishonest consumer n should be inspected and replaced before our proposed algorithm is re-invoked to obtain a more accurate regression analysis. However, the simulation results are omitted from the manuscript due to space constraints.



(a) 15 consumers

(b) 45 consumers

Fig. 6. Value of a and $(a + \beta)$ obtained by CVLR-ETDM when a is variable.

8. Conclusion

In this work, we have designed two algorithms, namely LR-ETDM and CVLR-ETDM, which are capable in identifying the dishonest consumers who are committing energy theft as well as locating the faulty equipment, with the aim to reduce non-technical losses due to energy thefts and metering defects in smart grids. The two algorithms are based on linear regression. Any non-zero anomaly coefficients are indicative of energy thefts or metering defects. We found that LR-ETDM might be unstable when there are inconsistent energy thefts and/or defective smart meters. Therefore, we incorporated categorical variables into linear regression and developed CVLR-ETDM so that the algorithm can successfully detect consumers' malfeasance and faulty meters even when there are inconsistent cheating trends/faulty equipment. Simulation results show that fraudulent consumers can be detected regardless of whether they steal energy at a constant and/or variable rate. As further work, we shall look into the noise tolerance issue of the proposed algorithms and design algorithms to conceal consumers' smart meter consumption data for preserving their privacy while still being able to identify the locations of malicious and defective smart meters.

Acknowledgment

Funding: This work was supported in part by the High Impact Research Grant (D000022-16001).

References

[1] Chauhan A, Rajvanshi S. Non-technical losses in power system: a review. In: Proceedings of 2013 international conference on power, energy and control, ICPEC 2013; 2013. p. 558–61.

[2] Han W, Xiao Y. NFD: a practical scheme to detect non-technical loss fraud in smart grid. In: IEEE ICC 2014-communication and information systems security symposium. p. 605–9.

[3] Navani JP, Sharma NK, Sapra S. Technical and non-technical losses in power system and its economic consequence in Indian economy. Int J Electron Comp Sci Eng 2012;1:757–61.

[4] Accenture. Achieving high performance with theft analytics; 2011.

[5] McDaniel P, McLaughlin S. Security and privacy challenges in the smart grid. IEEE Sec Privacy 2009;7:75–7.

[6] Faria L, Melo J, Padilha-Feltrin A. Spatial-temporal estimation for nontechnical losses. IEEE Trans Power Deliv 2015;8977. pp. 1-1.

[7] Rashed Mohassel R, Fung A, Mohammadi F, Raahemifar K. A survey on advanced metering infrastructure. Int J Electr Power Energy Syst 2014;63:473–84.

[8] McLaughlin S, Podkuiko D, McDaniel P. 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; 2010. p. 176–87.

[9] Li F, Qiao W, Sun H, Wan H, Wang J, Xia Y, et al. Smart transmission grid: vision and framework. IEEE Trans Smart Grid 2010;1:168–77.

[10] Wang W, Lu Z. Cyber security in the smart grid: survey and challenges. Comp Netw 2013;57:1344–71.

[11] Yan Y, Qian Y, Sharif H, Tipper D. A survey on smart grid communication infrastructures: motivations, requirements and challenges. IEEE Commun Surv Tut 2013;15:5–20.

[12] Xiao Z, Xiao Y, Du DHC. Exploring malicious meter inspection in neighborhood area smart grids. IEEE Trans Smart Grid 2013;4:214–26.

[13] Nizar AH, Dong ZY, Jalaluddin M, Raffles MJ. Load profiling method in detecting non-technical loss activities in a power utility. In: Proceedings of the first international power and energy conference (PECon 2006). p. 82–7.

[14] Iñigo Monedero RM, Biscarri Félix, León Carlos, Guerrero Juan I, Biscarri Jesús. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. Int J Electr Power Energy Syst 2012;34:90–8.

[15] McLaughlin S, Holbert B, Fawaz A, Berthier R, Zonouz S. A multi-sensor energy theft detection framework for advanced metering infrastructures. IEEE J Select Areas Commun 2013;31:1319–30.

[16] Xiao Z, Xiao Y, Du D-C. Non-repudiation in neighborhood area networks for smart grid. IEEE Commun Magaz 2013;51:18–26.

[17] Selvapriya C. Competent approach for inspecting electricity theft. Int J Innov Res Sci, Eng Technol 2014;3:1763–6.

[18] Khoo B, Cheng Y. Using RFID for anti-theft in a chinese electrical supply company: a cost-benefit analysis. In: Wireless telecommunications symposium (WTS); 2011. p. 1–6.

[19] Huang S-C, Lo Y-L, Lu C-N. Non-technical loss detection using state estimation and analysis of variance. IEEE Trans Power Syst 2013;28:2959–66.

[20] Jiang R, Lu R, Wang Y, Luo J, Shen C, Shen X. Energy-theft detection issues for advanced metering infrastructure in smart grid. Tsinghua Sci Technol 2014;19:105–20.

[21] Nizar AH, Zhao JH, Dong ZY. Customer information system data pre-processing with feature selection techniques for non-technical losses prediction in an electricity market. In: 2006 International conference on power system technology, POWERCON2006. p. 1–7.

[22] Mashima D, Cárdenas AA. 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; 2012. p. 210–29.

[23] Nagi J, Yap KS, Nagi F, Tiong SK, Koh SP, Ahmed SK. NTL detection of electricity theft and abnormalities for large power consumers in TNB Malaysia. In: Proceeding, 2010 IEEE student conference on research and development - engineering: innovation and beyond, SCOReD 2010. p. 202–6.

[24] Nagi J, Yap KS, Tiong SK, Ahmed SK, Mohamad M. Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans Power Deliv 2010;25:1162–71.

[25] Nagi J, Yap KS, Tiong SK, Ahmed SK, Mohammad AM. Detection of abnormalities and electricity theft using genetic support vector machines. In: Proceedings/TENCON of the IEEE region 10 annual international conference. p. 1–6.

[26] Depuru SSSR, Wang L, Devabhaktuni V, Green RC. High performance computing for detection of electricity theft. Int J Electr Power Energy Syst 2013;47:21–30.

[27] Jokar P, Arianpoo N, Leung VCM. Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans Smart Grid 2016;7:216–26.

[28] Salinas S, Li M, Li P. Privacy-preserving energy theft detection in smart grids: a P2P computing approach. IEEE J Select Area Commun/Suppl 2013;31:257–67.

[29] Salinas S, Li M, Li P. Privacy-preserving energy theft detection in smart grids. Annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks workshops, vol. 1. p. 605–13.

[30] Fang X, Misra S, Xue G, Yang D. Smart grid - the new and improved power grid: a survey. IEEE Commun Surv Tut 2012;14:944–80.

[31] Liu J, Xiao Y, Gao J. Achieving accountability in smart grid. IEEE Syst J 2014;8:493–508.

[32] Sahoo K, Nikovski S, Muso DN, Tsuru T. Electricity theft detection using smart meter data. In: Innovative smart grid technologies conference (ISGT). IEEE Power & Energy Society; 2015. p. 1–5.

[33] Amral N, Ozveren CS, King D. Short term load forecasting using multiple linear regression. In: 2007 42nd International universities power engineering conference. p. 1192–8.

[34] Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Arzteblatt Int 2010;107:776–82.

[35] Rodriguez G. Linear models for continuous data. Princeton Stat 2013.

[36] Studenmund AH. Using econometrics: a practical guide; 2006.

[37] Artes M. Statistical errors. Medicina clinica 1997;109:606–7.

[38] Chambers M, Dinsmore TW. Advanced analytics methodologies driving business value with analytics. 1st ed. Pearson Education, Inc.; 2014.

[39] Pedhazur EJ. Multiple regression in behavioral research, volume 3; 1997.

[40] Starkweather J. Categorical variables in regression: implementation and interpretation; 1997 <http://researchsupport.unt.edu/class/Jon/Benchmarks/CategoricalRegression_JDS_June2010.pdf> [accessed August 9, 2016].

[41] Skrivanek S. The use of dummy variables in regression analysis; 2009.

[42] Irish Social Science Data Archive (ISSDA); 2009 <http://www.ucd.ie/issda/data/commissionforenergyregulationcer/> [accessed August 9, 2016].

[43] Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comp Sci 1995;35:826–33.