2020-03

Using multiple data sources to detect manipulated electricity meter by an entropy-inspired metric

Hock, D

10.1016/j.segan.2019.100290
Sustainable Energy, Grids and Networks
Elsevier BV

Sustainable Energy, Grids and Networks 21 (2020) 100290

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

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

Using multiple data sources to detect manipulated electricity meter by an entropy-inspired metric

Denis Hock [a,*], Martin Kappes [a], Bogdan Ghita [b]

[a] Frankfurt University of Applied Sciences, Nibelungenplatz 1, Frankfurt am Main, DE, Germany
[b] University of Plymouth, Drake Circus, Plymouth, GB, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received 20 May 2019
Received in revised form 31 October 2019
Accepted 7 December 2019
Available online 18 December 2019

Keywords:
Electricity theft
Anomaly detection

ABSTRACT

With the digitalization of electricity meters many previously solved security problems, such as electricity theft, are reintroduced as IT related challenges which require modern detection schemes based on data analysis, machine learning and forecasting. Here, we demonstrate a multidimensional anomaly detection approach for the early detection of tampered with electricity meters by comparing a set of multiple energy demand time series. Our method can complement and enhance existing monitoring systems which usually only analyze a single time series. We aim to detect electricity theft, which leads to noticeable outliers in our work. We present three data preprocessing methods to produce outliers in case of energy theft and highlight the requirements and fine-tuning mechanisms for the aggregation and comparison of multiple data sources. We show that our metric is robust against multiple manipulated data sources, which is a concrete improvement to alternative outlier preserving concepts to aggregate multiple data sources. With detection rates better than 90%, we demonstrate the effectiveness of using several data sources simultaneously, that, when used individually, provide little value in anomaly detection. Furthermore, we show that we can use different households as comparable data sources, without clustering the households according to their similarity first.

1. Introduction

The success of renewable energy usage is fueling the power grids most significant transformation seen in decades, from a centrally controlled electricity supply towards an intelligent, decentralized power supply. However, as power grid components become more connected, they also become more vulnerable to cyber attacks, frauds, and software failures.

Many recent developments focus on cyber–physical security, such as physical tampering detection, as well as traditional information security solutions, such as the encryption of certain communication channels. However, information security cannot cover the entire challenge of cyber threats, as such digital meters can be vulnerable to software flaws and hardware malfunctions. Illera et al. [1], demonstrated that smart meters installed in Spain used strong symmetric encryption but stored a static encryption key in a plain text file, which allows adversaries to artificially manipulate and tamper with the data and measurements of a smart meters communication channel. For this reason, there is a high interest in utilizing the fine-grained data and recent advances in machine learning to detect electricity theft and data manipulation with data analysis and anomaly detection methods.

Here, we introduce such an anomaly detection scheme, which inspects the power measurements of a smart meter to detect electricity theft. In contrast to traditional anomaly detection approaches, which observe a single source over time to detect tampering, our approach consults the expected energy demand characteristics across similar data sources over time (in a matrix) and can unveil otherwise unseen outliers.

Due to the daily pattern of energy load curves, the detection of relevant outliers by comparing different data sources (e.g. electricity meters) or different data sets of one data source (e.g. previous days) can be a challenging task. Our contribution is to illustrate an entropy-inspired outlier preserving metric which can be used to aggregate data, so that repeating patterns can be removed. We showcase two scenarios to tamper with electricity meter data. Furthermore, we introduce and validate requirements on the input data and highlight cases in which our entropy-inspired metric leads to a concrete improvement compared to alternative methods aggregating different data sources.

In the remainder of this paper, we first present an overview of our proposed framework and discuss three alternative input data, derived from raw data, as well as the benefits and limitations of using multiple data sources and our metric. Our experimental validation showcases the manual optimization and fine-tuning

* Corresponding author.
E-mail addresses: dehock@fb2.fra-uas.de (D. Hock), kappes@fb2.fra-uas.de (M. Kappes), bogdan.ghita@plymouth.ac.uk (B. Ghita).

https://doi.org/10.1016/j.segan.2019.100290
2352-4677/

of input data and an extensive evaluation of parameters that influence the detection rate of our approach. We compare the feasibility of our method which removes repeated pattern and an alternative method which does not consider these patterns. Furthermore, we demonstrate the advantage of our entropy-inspired method with the juxtaposition of an alternative outlier preserving aggregation method. In our final experiment, we compare our approach with two alternative anomaly detection schemes based on Naive Bayes and XMR charts.

2. Related work

Many European countries launched a conversion to the next generation power grid to fully benefit from and address the challenges associated with distributed energy generation. The wide availability of high resolution electricity data at residential level and rapid advancements in machine learning techniques, brought in a number of related research questions ranging from energy forecast to typical load profile classification, to support the safety critical processes in the power grid. In the following, we summarize the relationship between these topics, to smart grid security and anomaly detection.

2.1. Smart grid

Early studies on energy demand, e.g. Bohi et al. [2], often focus on energy forecast and corresponding modeling methods with various inputs, such as weather data and socioeconomic data. A new trend is to extract such data from energy demand to monitor security and safety aspects of the smart grid. Previous work go as far as identifying individual appliances, socioeconomic data and personal behavior with the varying energy consumption of electric appliances. E.g. Kleiminger et al. [3] used multivariate methods and supervised learning to detect human presence. Molina-Markham et al. [4] used density-based clustering and supervised learning to identify private information about consumers. Yohanis et al. [5] analyzed the effect of the number of occupants and the size of dwellings on load curves. Price [6], as well as Druckman et al. [7] were able to find income and employment status of households. Carroll et al. [8] and Mcloughlin et al. [9] correlated load curves to employment status and presence of children. Kolter et al. [10] analyzed the relation between demand and building properties, such as the number of rooms and the building value. Beckel et al. [11] extracted the number of occupants. Newing et al. [12] associated energy consumption patterns with particular dwellings, income and number of children.

2.2. Energy theft

Smart Grids are historically not designed with Internet security in mind, as mentioned by Jain et al. [13], but security flaws can result in customer information leakage and a cascade of inadvertent or deliberate failures, such as a massive blackout and destruction of infrastructures as introduced by Metke et al. [14]. Recently, Westerhof [15] simulated the disastrous consequences of a coordinated cyber-attack on photovoltaic systems, which may lead to a national power outage. In a study of Dabrowski et al. [16], IoT botnets use common devices, connected to the Internet to selectively increase and decrease power consumption which can lead to falling below the standard frequency and ultimately to power outages. These reasons advanced the topic to a rapidly growing research area, as several surveys [17–19] show.

A particular active and challenging field in the area of smart grid security is electricity theft.

The basic function of a smart meter is the measurement of the energy consumption for billing, which is calculated by adding up the mains voltage multiplied with the current drawn by active devices. We can split the methods to tamper with meters in intrusive methods inside the meter housing and non-intrusive methods outside the meter. Common intrusive methods include attaching electrically conductive objects to pass current away from the measurement circuit, disconnecting the phase to interrupt the measurement or exchanging the phase connection to archive a negative measurement. Non-intrusive measurements include the usage of strong magnets to temporary disable the power supply of the meter. For this reason, billing companies are interested in detecting energy theft by leveraging the fine-grained smart meter data and machine learning. Many authors [20–22] employed machine learning to classify consumption pattern and load profiles in order to detect electricity theft. Cardenas et al. [23] developed a game theoretic scheme between adversary and billing company. Bandim et al. [24] proposed a central observer to compare the total energy consumption with the reported consumption of individuals. Salinas et al. [25] suggested a distributed algorithm to compute the trustworthiness of each participant. Spiric et al. [26] detected energy theft by monitoring the energy consumption with XMR charts. In addition to the physical methods mentioned above, modern smart meters pose the additional danger of sophisticated digital manipulation methods, which are not covered in this paper. With digital access to measurements and metering information as well as knowledge on the detection method, the adversary could potentially aim for stealthy manipulation scenarios such as mimicry attacks. However, such methods are general limitations of anomaly detection systems and not specific to energy theft, as introduced by Urbina et al. [27] and Bouche et al. [28].

2.3. Anomaly detection

A potential method to unveil energy theft is anomaly detection. Anomalies are defined as deviations from the expected data, rather than by predefined malicious data. Therefore, it is particularly suitable to detect previously unknown tampered with data, which is difficult to describe in the volatile and highly heterogeneous energy demand.

Since the initial scientific publication of Denning [29], the popularity of anomaly detection in order to detect malicious behavior is constantly increasing. The emergence of sensors with processing and communication capabilities stimulated great interests in anomaly detection on smart grid components. Several surveys [30,31] show approaches for anomaly detection on sensor networks. A few authors also focus on components of the power grid, e.g. Braun et al. [32] used the minimum co-variance determinant to detect faults in photo-voltaic arrays and Dienst et al. [33] consults change-point analysis to observe the condition of photo-voltaic power plants.

Similar to our approach, Mclaughlin et al. [34] developed the anomaly detection scheme AMIDS, based on network data and electricity data, in order to detect energy theft. AMIDS utilizes electricity measurements together with a NIALM database to label the amplitude changes in a time series and subsequently learns them with the Naive Bayes algorithm.

Furthermore, Raciti et al. [35] explored smart meters embedded with anomaly detection to identify threats on cyber physical systems. Rossi et al. [36] proposed to take collective and contextual anomalies into account to detect events such as over-voltages and under-voltages. Mookiah et al. [37] introduced a graph-based anomaly detection approach, where vertices represent smart appliances and edges represent their usage, to detect anomalies

in power usage. Yip et al. [38] presented an anomaly detection scheme that adopts linear programming to detect energy theft and reduce false positives by taking into consideration the impact of technical losses and measurement noise. Fengming et al. [39] detected anomalies, such as short circuit faults, by comparing a time-series reconstructed by a recurrent neural network with the original data. Andrysiak et al. [40] presented a solution to detect energy theft with network traffic anomaly detection in critical smart metering infrastructure. Zhou et al. [41] aimed to detect outliers such as communication failures and voltage disturbances by comparing multiple time series of voltage with a randomized block coordinate descent algorithm.

Our anomaly detection approach utilizes an entropy-inspired metric to model the expected energy demand of different data sources. Entropy-inspired anomaly detection have been applied in other areas such as healthcare by Richman et al. [42], biodiversity assessment by Vranken et al. [43], or network anomaly detection by Wagner et al. [44], but not as a method to aggregate different data sources.

3. Data and method

Multiple data sources or multiple data sets of one data source are a valuable extension to an anomaly detection model as the comparison to similar data can point out otherwise hidden outliers. However, even the differences between data sources often show regular patterns and hence, manipulated electricity data often leads to local outliers instead of easy to find global outliers. The aggregation of data to a single time series is essential to remove these patterns and can significantly improve the detection rate.

For our approach, we first define a data preprocessing method which results in a less noisy and standardized time series of values, which reflects the activity of electric appliances found in the energy demand load curve and produces outliers in case of tampering. We refer to these values, which are the input data for the aggregation method, as 'feature' to separate them from the 'metric', which are the output of the aggregation method.

These features build the underlying statistical model and distinguish between regular and anomalous behavior. We introduce three features, each aiming to characterize the activity of a load curve. In the next step, each feature is analyzed with regard to a comparable data, such as a previous day of the same household or a household with similar energy consumption: the features are aggregated by encoding their distribution to a normalized floating point value with our so called entropy-inspired metric, which preserves outliers in the distribution of input features. The resulting time series can be combined with off-the-shelf forecast algorithms, such as Holt–Winters, to remove the daily patterns by subtracting the forecast. Now, a simple threshold can distinguish benign and anomalous values. Fig. 1 shows the complete process to build the normal model and subsequently detect anomalous time windows. The left part illustrates the data at each step of the process: the top shows the raw input of different sources as line chart, the middle shows the derived features, whereas each time window results in a stacked bar of all data sources, the bottom shows the resulting entropy-inspired metric. The right of our figure shows training and anomaly detection process with corresponding parameter. While the computation of the metric is equal for both, as last step the forecast model is subtracted from the actual metric to remove the daily pattern. Now, a simple threshold can classify benign and manipulated time windows.

In the following, we provide a comprehensive overview of all steps in the method and discuss on their benefit, starting with our choice of raw data and the derived features.

3.1. Features

The output of the data preprocessing produces an outlier in case of manipulated electricity meter. As these data are intermediate results and input for our entropy-inspired aggregation method we call them *feature* to keep them apart from the output of the aggregation, which is in the following called *metric*.

The efficient operation of the power grid depends critically on monitoring the participants, which is accomplished by using measurements collected from meters deployed throughout the grid. Typically, measurements include the real, reactive, and apparent power as well as phase volt measurements and current. Monitoring the current and voltage is critical for many applications such as the fault monitoring and the early detection of over-voltage or power line failures. The reactive and apparent power need to be closely monitored and counterbalanced by power grid operators to avoid unnecessary thermal line losses. The mains frequency, which measures the balance of production and consumption, helps to prevent overproduction that destroys equipment and underproduction that can lead to blackouts. From this perspective, the manipulation of each measurement can lead to safety critical situations. However, in this work we focus on the integrity of real power measurements, because the tampering with real power, which depicts the amount of work performed by an component, is easy and intuitive for an adversary attempting billing fraud. Furthermore, the manipulation of power measurements can be used as foundation for many other sophisticated scenarios, such as a forged blackout and other scenarios requiring multiple corrupted electricity meters, which are not in the scope of this work.

A simplistic and straightforward approach to detect an unexpected energy demand is to classify the raw data measurement by measurement. However, raw load curves are difficult to compare because the aggregated and overlaying patterns of several components can drastically change and pollute the appearance of a load curve. To clearly distinguish legitimate and anomalous data, we need to filter any unnecessary information, which would produce the same output for different inputs, and derive features which are less noisy and more resilient to changes. The regular activities of a household, given by e.g. the number of residents, work hours, and sleeping habits are easy to find and still more consistent than raw data. Therefore, we propose to numerically characterize the 'activity' of a time period in a households energy demand. We can see activity as an occurrence of a state change for one or many appliances and quantify the number of visible operations. Or we can see activity simply as the consumption above a certain level, which excludes stand-by devices. As we neither need to distinguish devices nor find the cause of any activity in order to perform anomaly detection, a consistent behavior of the features is sufficient for our purpose. We implement three features according to the above criteria, which all estimate the amount of state changes or consumption during a predefined time window to summarize the activity: the number of high amplitude points (f_1), the number of amplitude changes (f_2) and the number of similar amplitudes in a row (f_3). All three features are computed by applying a binary classification on each measurement and adding up the number of measurements with positive class. A lower activity results in a higher score and hence an outlier in case of a manipulated electricity meter. By counting the number of measurements in a time window we always receive results with a fixed range, which are easy to compare and to normalize.

Remark 1. Formally, consider a finite time series $\mathcal{T} = x_1, x_2, \ldots, x_n,\ x_i \in \mathbb{R}_0^+$ *for all* $1 \leq i \leq n$, with n elements representing energy demand. Then, each feature is a conditional sum over $\mathcal{T}$, formally a function $f : \mathcal{T} \mapsto \mathbb{N}_0$ with a range $[0, n]$, here

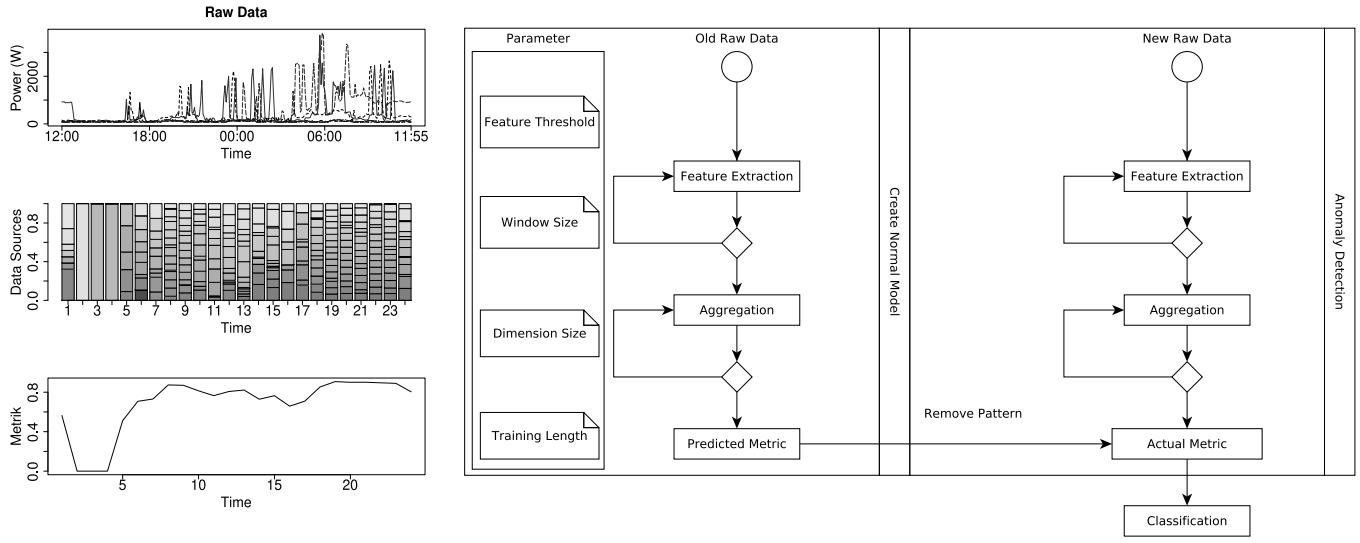


Fig. 1. Workflow of the anomaly detection theme.

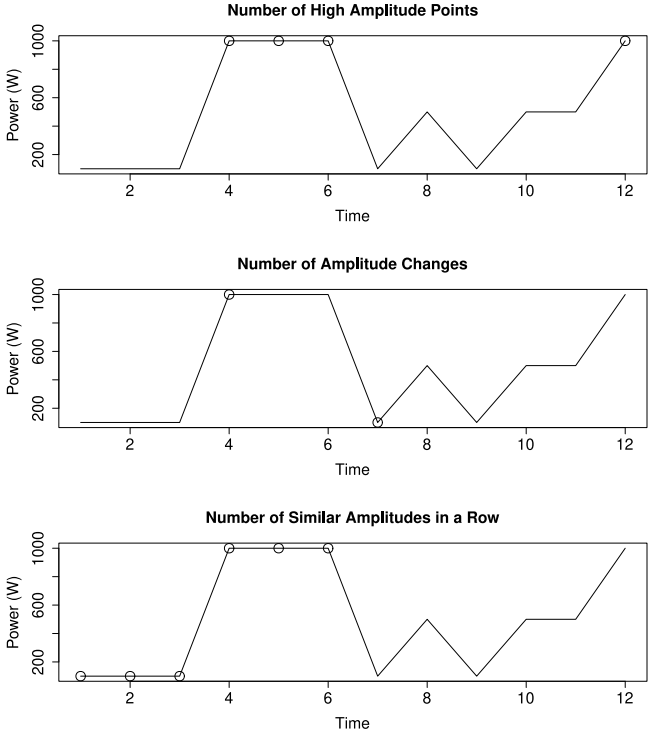


Fig. 2. Illustration of our feature extraction f_1(top), f_2(middle), f_3(bottom).

represented as Iverson bracket with ε defined as threshold in Watt:

$$f_1(\mathcal{T}) = \sum_{i=1}^{n} [x_i < \varepsilon] \tag{1a}$$

$$f_2(\mathcal{T}) = \sum_{i=1}^{n} [x_i - x_{i+1} < \varepsilon] \tag{1b}$$

$$f_3(\mathcal{T}) = \sum_{i=1}^{n} [l(x_i, \varepsilon) > \delta] \tag{1c}$$

The function $l(x_i, \varepsilon)$ returns the amount of neighboring measurements in a row, that are equal to x_i, if we round by ε, whereas δ is a threshold for the number of neighbors. We define f_1 and f_2 as *less than*, because we want a big number in case of energy theft. However, this is only more intuitive for us and also works the opposite way.

Fig. 2 illustrates our methods to extract features. The line shows an artificial energy demand (y-axis) over time (x-axis), whereas the points illustrate measurements in one of two classes for 'High Amplitude' (top), 'Amplitude Change' (middle) or 'Similar in a Row' (bottom).

3.2. Dimensions

In the following, the most important characteristic of the features is to provide distinctive and discriminating distributions. Here, we point out how to efficiently utilize such a distribution for anomaly detection.

Intuitively, energy demand is highly adaptive and ever changing. Due to this fact, a static absolute value or maximum difference that unveils tampering is difficult to design. We can mitigate this concept drift by using a similar data, such as historic data or spatially close and structurally identical components that naturally adapt due to the similar conditions. We call these data sources or data sets of a data source *dimensions*, because the regular measurements of a smart meter are mathematically considered a finite time series $\mathcal{T}$ (as defined before). As such, the time synchronous results of m smart meters are a $n \times m$ matrix $\mathcal{M}$. *Note that, a dimension, in our definition, does not necessarily correspond to a physical dimension, but anything that can construct a matrix of comparable data (e.g. it would be possible to add a dimension for: all single-family households, all households of a region, or all Saturdays).* For our experiments, we address a third dimension by subdividing the time-axis of different meters in k periods of time (1 day), which is useful because we expect the same pattern over these periods. As a result, the dimensions in our matrix compare:

(a) Time: n measurements for adjacent times (traditional)
(b) Households: measurement at time i for m smart meters
(c) Dates: measurement at time i for k different periods of time

To demonstrate the utility of different dimensions, we can use an example where a single dimension is insufficient for detection. In this context, we can define the concept of anomalous as an outlier different from the majority of compared data and assume a $n \times m \times k$ matrix $\mathcal{M}$, with identical values x, with some of these

values being manipulated through division (identified in gray in the equation below)[1]:

$$\mathcal{M} = \left[\begin{bmatrix} x_{111} & \cdots & x_{1m1} \\ \vdots & \ddots & \vdots \\ x_{n11} & \cdots & x_{nm1} \end{bmatrix} \begin{bmatrix} x_{112} & \cdots & x_{1m2} \\ \vdots & \ddots & \vdots \\ x_{n12} & \cdots & x_{nm2} \end{bmatrix} \begin{bmatrix} x_{11k} & \cdots & x_{1mk} \\ \vdots & \ddots & \vdots \\ x_{n1k} & \cdots & x_{nmk} \end{bmatrix}\right] \tag{2}$$

By aiming to detect outliers different from the majority, we can see that the result greatly depends upon the dimension: comparing $x_{111}, x_{211}, \ldots, x_{n11}$ (in dimension 'Time') results in no detection because all values are equally manipulated. Comparing $x_{111}, x_{121}, \ldots, x_{1m1}$ (in dimension 'Date'), half of the values are manipulated and we cannot distinguish normal and manipulated values. Comparing $x_{111}, x_{112}, \ldots, x_{11k}$ (in dimension 'Households') we can clearly see an outlier, because only x_{111} is different from all other values. We can simplify the relation of dimensions and detection rate by considering each dimension of the matrix as individual anomaly detection approach. By assuming that each dimension is statistically independent and used for classification simultaneously, we can use the binomial formula $P = \binom{n}{k} \cdot p^k (1-p)^{n-k}$ to compute the total probability for detection. Whereas P computes the cumulative probability that the anomalous value is k times successfully detected in a total of n dimensions, if each dimension has the same individual detection rate of p. Fig. 3 visualizes the probability that at least one dimension detects the manipulated data. Each line shows a different probability for a correct classification of the individual dimensions, whereas the x-axis shows the amount of trials. The y-axis shows the cumulative probability for at least one correct classification.

3.3. Entropy

The entropy is a convenient way to aggregate several values without losing information on outliers. We aim to aggregate the distribution of each dimension to a single value, so that we can apply time series prediction and hence remove predictable patterns. Our reasoning here is that similar data sources, such as two spatially close and structurally identical photovoltaic cells or two days of the same household, should result in similar volumes of features, and hence a uniform distribution (or at least a consistent pattern). A sudden shift of the distribution due to a significant change of measurements is, according to this reasoning, an outlier and unexpected.

Shannon [45] defined the entropy as 'uncertainty', which is maximized when the outcomes of a random variable are equally likely. A frequency distribution of equal outcomes corresponds to a uniform distribution.

Remark 2. Formally, let $\mathcal{X} = x_1, x_2, \ldots, x_n$, denote the frequencies of outcomes from the random variable. Whereas $m = \sum_{i=1}^{n} x_i$ implies the number of all experiments. Then, the entropy $H(\mathcal{X})$ is defined by

$$H(\mathcal{X}) = -\sum_{i=1}^{n} \frac{x_i}{m} \cdot \log_2\left(\frac{x_i}{m}\right) \tag{3}$$

[1] We simplified the matrix notation of each row and column $x_{111}, x_{121}, \cdots, x_{1m1}$ for space considerations.

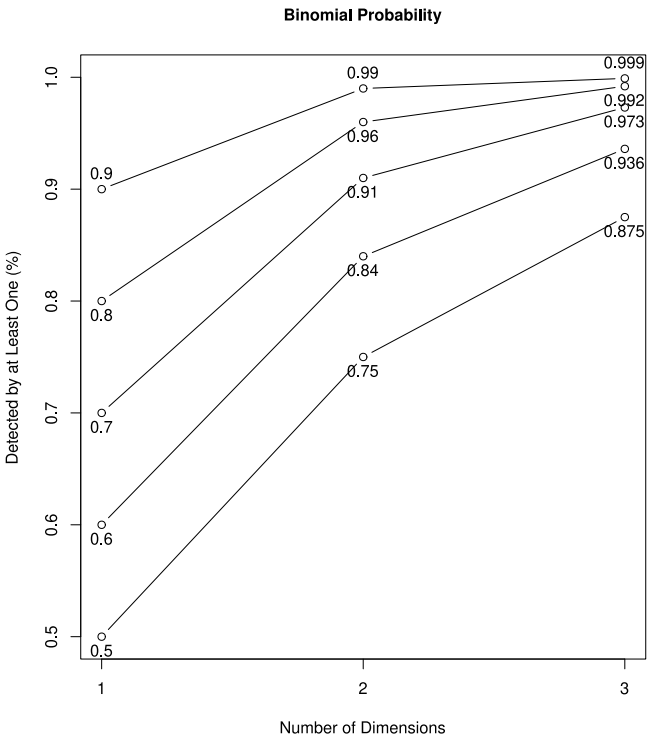


Fig. 3. Binomial probability for different dimensions.

If we define $0 \cdot \log_2(0)$ as zero, the entropy lies in a range of $[0, \log_2(n)]$, minimized if a single frequency differs from all others and maximized if all occur equally often. For a better comparison, we normalize the entropy to [0, 1] as follows:

$$\hat{H}(\mathcal{X}) = -\sum_{i=1}^{n} \frac{\frac{x_i}{m} \cdot \log\left(\frac{x_i}{m}\right)}{\log(n)} \tag{4}$$

We use the entropy not in a traditional sense but as a metric to mathematically trace outliers in the distribution of features over time. While our entropy-inspired metric may have characteristics similar to Shannon's entropy, we do not aim to proof that the entropy is the only function fulfilling the requirement to preserve outliers. The practical function as a metric, which encodes distributions without losing information on outliers, is the most important aspect for us.

The input for the entropy could principally be an artificial histogram, such as the relative volume of energy demand. However, if the range of input values changes over time, the entropy is difficult to compare with a previously computed entropy. Therefore, the volume of each feature should have a fixed range. Furthermore, the entropy is profoundly affected by the number of possible outcomes, which means more input values (dimensions) result in the entropy-inspired metric being less affected by outliers.

Let us assume that the initial state is a uniform distribution of a vector $v = (1/n, 1/n, \ldots, 1/n)$ of length n. It is reasonable that a change (a) to $v = (0, 0, 0, 1)$ is more visible than a change (b) to $v = (1/n-1, 1/n-1, \ldots, 1/n-1, 0)$, because the sum of changes in (a) is $2(n-1)/n$, whereas the sum of changes in (b) is $2/n$. According to this reasoning, we can construct our method to detect one of two situations using the entropy metric: detect few falsified value which must be much bigger than the normal values, or detect a majority of falsified values which must be much smaller than normal values. Here, we prefer the first option in order to detect anomalies early on and tolerate that visibility decreases with increasing proportion of tampered values.

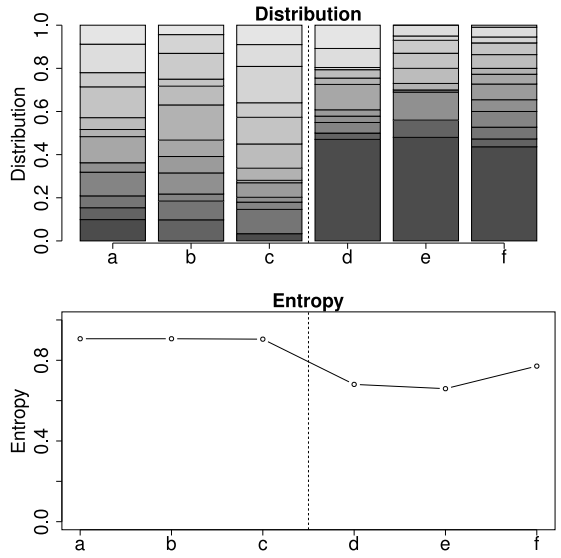


Fig. 4. Entropy values from uniform distribution to heavily skewed distribution.

The Fig. 4 illustrates the entropy of a distribution from real world data. On the *x*-axis of the plot, we can see six time windows: (a) morning, (b) afternoon and (c) evening of sample one, as well as (d) morning, (e) afternoon and (f) evening of sample two with energy theft. In the upper stacked bar plot, each stack (a–f) has twelve elements. Each element represents the size of our feature (f_1: Number of high amplitudes) for a different day during the corresponding time window. The left side (a, b, c) with normal energy demand is roughly uniformly distributed, whereas the right side (d, e, f) with energy theft includes an outlier. The bottom line diagram shows the corresponding entropy, which is high for a uniform distribution and low for a skewed distribution as in case of energy theft.

With the above characteristics of our metric, an anomaly is revealed by very small entropy values. However, such anomalies can also be detected by the outlier in the distribution of input values. Hence, it is always possible to detect the same anomaly by looking at the input values instead of the entropy. The entropy is only a convenient way to aggregate several values without losing this information (e.g. compare several households simultaneously over time). The main motivation to aggregate the results is to simultaneously apply off-the-shelf time series algorithms such as Holt–Winters on several households or days, which is difficult with a matrix representation. To apply the entropy horizontally on the dimension 'time' is not suitable as it would mean to encode an outlier which is already visible with other close values and worsen the result.

4. Experimental evaluation

Load curves often mirror specifics of the power grid, which can be hard to mathematically trace but for human operators with expert knowledge easy to estimate (e.g. the expected effect of energy theft, the normal difference of two regular load curves at peak times or the probability of legitimate outliers on specific days). In order to dynamically adjust the parameters according to such specifics, we introduce some relevant statistical indicators from a holistic data analytic viewpoint (without considering the mechanics of the power grid, which are well known to operators). It is possible to mimic these manual tuning rules for automation, but the efficient automatized tuning, which is a topic in the realm of optimization algorithms, is not within the scope of this paper.

In the following, we start with the introduction of our experimental data set and tampering methods. Then, we evaluate the significance of our features over raw data and especially consider the influence of our parameters to generate outliers in case of electricity theft. We provide a in-depth analysis on how the threshold ε, the window size, the number of dimensions and the length of training data for a forecast algorithm affects our result.

4.1. Data set and energy data tampering

In the recent past, several data sets, monitoring household electricity and environmental parameters have been released publicly. Researchers used these data set to prove the validity of their work for real life settings. This study uses the Electricity Consumption & Occupancy (ECO) data set provided by Beckel et al. [46], which is a comprehensive data set for non-intrusive load monitoring and occupancy detection research, offering individual appliance and occupancy readings every second. The data set offers real world power measurements of six houses over, depending on the household, a period ranging from four to eight month. The real power (W) values of smart meters are based on the SML-protocol, which captures the mean-cycle-power. The ECO data set provides, apart from declared exceptions, measurements in Watts with 4 decimal places for a total length of one year and for six different houses. For the purpose of this study, measurements were aggregated with a resolution of five minutes, as the fine grained resolution is neither necessary nor practically feasible for real world setups. In this data resolution, the raw values of the ECO data set are used as input data to compute the features 'number of high amplitude points' (f_1), 'number of amplitude changes' (f_2) or 'number of similar amplitudes in a row' (f_3) for a time window of predefined length. This process is repeated on each dimension, so that the different data sources can be aggregated to the metric.

Note that in the following our experiments for dimension 'Date', always shows the ECO data of household 1 (June 2012 to January 2013), while experiments for dimension 'House' utilizes the data of all six households. In case of missing data we removed the corresponding day from all households, which resulted in approximately 120 days which are simultaneously available for all households. For our experimental setup, we also utilize the real world measurements from the ECO data set to artificially construct cases of tampered with data. In order to motivate a realistic scenario, we manipulate the data according to the traditional (physical) tampering methods introduced in the related work section. Namely, aiming to bypass energy consumption and slow down the measurement or stop the measured energy consumption altogether. Such patterns are simple, but realistic scenarios. We assume that the adversary cannot gain unlimited digital access to the smart meter, and hence stealthy energy theft attempts, such as mimicry attacks, which attempt to bypass anomaly detection, are not in our scope.

For following experiments, we use two types of falsified demand with different influence on the original data. *Type 1* is the original data with an arbitrary region, with a length of at least the time window for a feature, replaced by 0 Watt and represents a case where the smart meter is cut off for a certain time, whereas *type 2* is the original data divided by 5 and represents a case where the smart meter is continuously manipulated to lower the demand. The type 1 falsified data may sound statistically trivial to detect, but regions without power also occur in legitimate load curves (e.g. sleeping hours or working hours) and hence only change the regular load patterns to a decreased activity.

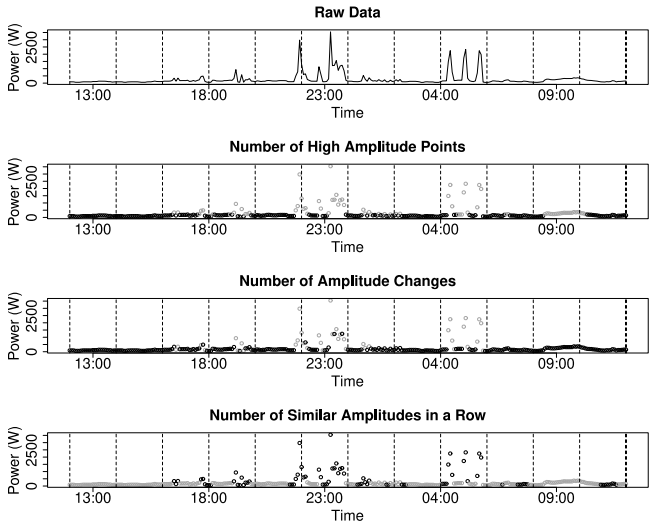


Fig. 5. Features with $\varepsilon = 200$ W.

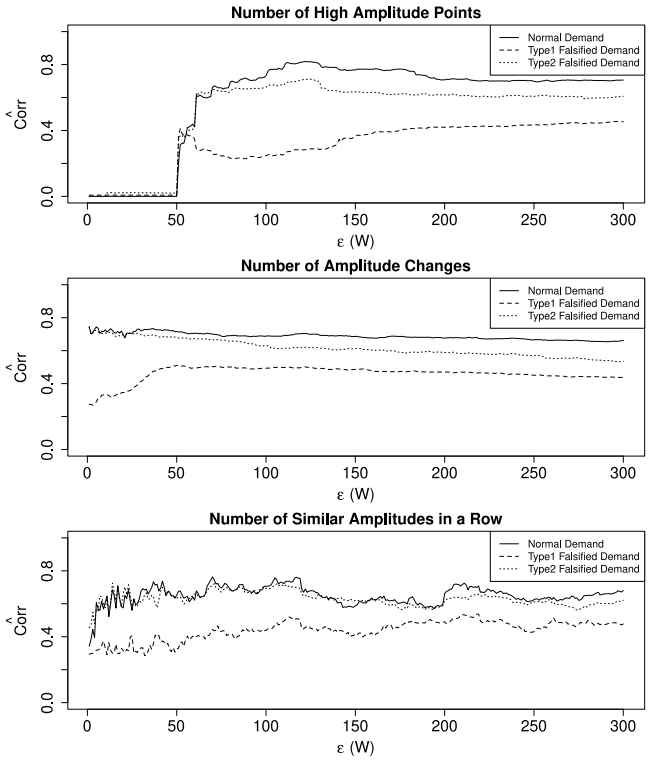


Fig. 6. Correlation of normal and anomalous load curves with ε.

4.2. Parameter evaluation: Features

Here, we optimize ε, the threshold parameter defining which measurements are classified as 'activity' for each feature. The threshold can be used to define which deviations from the normal model of a load curve are still within our expectations and must be found for each individual electricity meter. We start to evaluate the correlation of our features for benign and tampered load curves using different ε and subsequently introduce additional indicators on the detection quality of the features. Fig. 5 illustrates our features according to Section 3.1. The top plot 'raw data' shows one day of energy demand from the ECO data set, the following plots present the classification of measurements according to the condition of Eqs. (1a), (1b), (1c). The vertical dotted lines visualize the length n of a period which is used to compute the sum over all gray dots.

We proceed to assess the quality of ε in several steps: we want to ensure that our features can distinguish anomalous and normal data and can describe the relationship between two data sources. Here, we measure this property of our features using the *correlation*. Furthermore, as each feature is basically a classification of individual measurements, we show whether the measurements are evenly distributed over both classes, because a single-sided classification does not contain any information. On top of that, we illustrate whether measurements of different normal load curves are equally classified or random, depending on ε. In the following, we call these two properties *regularity* and *certainty*.

Our first objective is to see the influence of ε on the correlation, which shows the amount of variation that cannot be explained when the features of two load curves are compared. For this experiment, we sliced the energy demand of a single day to equally sized time windows (see Fig. 5) in order to compute our features, which results in a vector of features $\vec{v}$ representing energy demand.

Remark 3. A correlation of 1 indicates that time series $\vec{v}_1$ reacts at any time exactly like time series $\vec{v}_2$, while a correlation close to 0 means that the two curves are not related. A negative correlation shows that $\vec{v}_1$ and $\vec{v}_2$ is horizontally or vertically reversed. We are not interested in negative correlation, which would roughly mean that $\vec{v}_1$ shows energy production while $\vec{v}_2$ shows energy consumption or that $\vec{v}_1$ shows high consume in the morning while $\vec{v}_2$ shows high consumes in the evening. For this reason, we change the scale of the correlation (in Fig. 6) to [0, 1], whereas 0 means different and 1 similar as follows: $\hat{corr} = {}^{(corr(\vec{v})+1)}/_{2}$.

Here, we are interested on whether the features of a normal load curve show higher correlation to the features of a normal load curve than to a falsified load curve and evaluate the correlation to falsified demand with different load pattern (type 1: region replaced by 0 Watt) and decreased energy (type 2: region divided by 5).

Fig. 6 presents the normalized correlation (y-axis) with different ε (x-axis) for each feature (avg. 30 days). The black line plots the $\hat{corr}$ of normal data, and should be maximized, whereas the dotted lines each represent the comparison to tampered data, and should be minimized. The plot 'Similar in a Row' is not expected to show a minimized correlation for type 2, because the pattern, and hence measurements in a row, is unchanged through the tampering method. In 'High Amplitude Points', we can see that $\varepsilon < 50$ W results in a $\hat{corr}$ of 0 while both other features work with a small ε. This is of course coined to this specific household, which consumes (due to the standby power of many appliances) at least 50 Watt even without human activity. Next, we introduce the computation of *regularity* and *certainty* for Fig. 7.

Remark 4. Naturally, $R = {}^{f_1(\mathcal{T})}/_{n}$ shows the ratio between {0, 1} classifications. Let us define regularity as ratio approaching 0.5, which can be normalized to range [0, 1], where 0 is one-sided and 1 regular with $Regularity = 1 - {}^{|R-0.5|}/_{0.5}$. Using the *certainty*, we want to ensure that the classification is not random. Here, we use the equation for our feature without the sum, which results a binary time series $\mathcal{T}' = b_1, b_2, \ldots, b_n,\ b_i \in \mathbb{B}$ *forall* $1 \leq i \leq n$ representing one day of energy demand. Let us define a matrix M, where each of the m column is an instance of T' showing a different day, then a row is 'random', if the {0, 1}-ratio of the m instances is 0.5. We define certainty as the opposite of random,

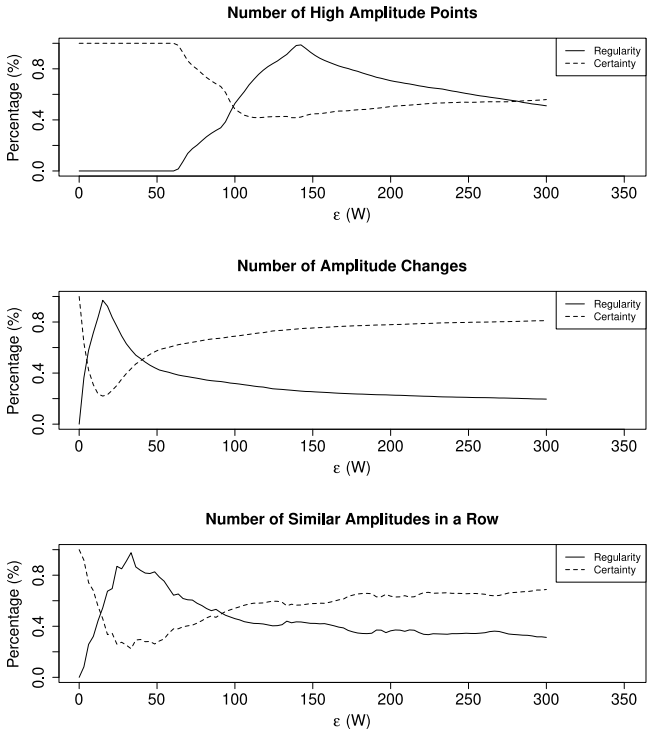


Fig. 7. Regularity and certainty with regards to ε.

$Certainty = {}^{|RowRatio-0.5|}/_{0.5}$ with a range [0, 1], where 0 means random and 1 means that b_i is equal for all columns.

In Fig. 7 (mean of 30 days), we show the relationship of *regularity* and *certainty* (y-axis) for each feature. The x-axis shows different thresholds ε in Watt, which are used to adjust the conditional expression. As we only use normal data in this experiment we expect that a good threshold would maximize the certainty. A certainty of 1 would mean that the pattern of all 30 days is exactly the same, which is good for anomaly detection because a different load pattern would stand out. The regularity is only used to identify errors, e.g. thresholds with a one-sided classification which cannot distinguish between different days at all and for this reason also results in a high certainty. For the anomaly detection itself a low regularity would be perfectly fine. According to this reasoning, we can see that the certainty starts with a high value, but as the regularity at this point is zero, it only means that the feature is unable to distinguish different load curves. Next, we can see a negative peak for the certainty (approximately for the same ε that maximizes the regularity), which should be avoided, as zero certainty means the classification of all 30 days was random. The two intersections of regularity and certainty are a good compromise and show us where the explanatory power of the feature is maximized. The result also confirms the estimation, of the previous figure showing the correlation of normal and anomalous load curves, which showed that 'Number of High Amplitudes' needs a high ε (for this household) while both other features work with low thresholds. New to Fig. 7 is, that we can now see that the feature 'Number of Amplitude Changes' results in an overall higher certainty, which means that the patterns of this feature are more stable and therefore easier to forecast than the other features.

Altogether, we conclude that ε can be adjusted, so that the features react well to specific tampering methods or, if the tampering method is unknown, to the energy demand of a certain data source.

4.3. Parameter evaluation: Entropy

In our experiments, we demonstrated that, by using our features, we can capture unexpected energy demand changes and that the unexpected values appear as outlier in contrast to normal data. We can maximize the change of the entropy using two parameters: first, the window size of our feature, which is defined by the amount of measurements in a window, and affects the maximum difference of normal data and outlier. Second, the size of the vector, which is defined by the number of dimensions, and affects the proportion of normal and falsified values in the distribution. Here, we examine these two parameter in conjunction with the entropy.

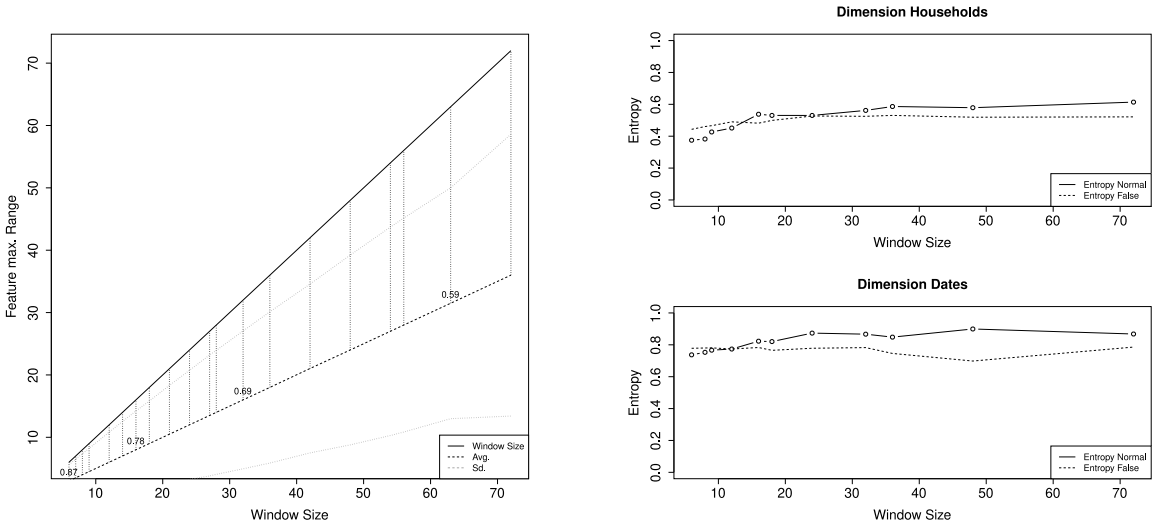
The window size, which is used to compute a feature, defines the amount of classified measurements and hence the range of the feature. In Fig. 8(a), we show the window size (x-axis) and corresponding avg. feature volume (y-axis) of normal data (mean of 24 days). While the avg. volume is consistent about half of the window size, we can see that the standard deviation slowly decreases with the window size (total range) of the feature. However, a bigger window size can be a disadvantage, because it will only appear as anomalous if the majority of measurements within this window are falsified. Fig. 8(b), shows the average entropy for six normal samples (black) and one of six samples falsified (dotted) for dimension 'household' and 'days'. *Note that, in our paper falsified data should always result in an outlier and hence, a small entropy.* We can see here, that the distance of the entropy with normal and anomalous data is increasing with the window size of the feature. The avg. entropy is not expected to be a particular good indicator, because not every time window is expected to be uniformly distributed, and hence the avg. entropy is influenced by regular daily pattern. However, the fact that small window size can result in a completely wrong model, where the avg. entropy of tampered data is greater than the entropy of normal data, indicate that the entropy of such time windows may be hard to predict.

The second parameter we need to address is the 'dimension size', which is of paramount importance to our entropy-inspired metric, because it reflects the necessary distance to other (normal) values in order to appear as outlier. In Section 3.3, we introduced the theoretical concept of the dimensions and demonstrated the effect of energy theft (outlier), which resulted in a smaller entropy. It may sound intuitive that smaller dimension sizes are better, because in a vector of smaller length, individual false values appear proportionally bigger and result in a more skewed distribution which affects the entropy. But smaller vectors are often heavily influenced by the standard deviation and appear anomalous even without falsified data.

Fig. 9 illustrates this effect by the distribution (bar plot) and entropy (line chart) of our feature over different days in 4 h windows (48 measurements per window) for two different dimension sizes (top: 4 days; bottom: 14 days). Similar to Fig. 4 in the introduction of entropy, each stack of the bar plot shows the distribution of our feature across several days. If one element of the stack is bigger than all others (as in case of energy theft), this outlier results in a smaller entropy.

The left figures show regular days while the middle figures show a sample with one day falsified (type 1). In the upper figures, we can see the potential problem with a too small dimension size: Due to the high standard deviation, some time windows of the normal data include legitimate outliers (e.g. left upper bar plot - stacked bar nr. 5). These legitimate outliers cannot be distinguished from energy theft. *Note that, we had to look for a specific day with irregular load curve in our data set to demonstrate this.*

According to this reasoning, the optimal dimension size is the smallest possible size, where we can still consistently distinguish



(a) Entropy for different window sizes.

(b) Entropy for normal values only and 1 of 6 samples falsified.

Fig. 8. Effect of the feature window size on the entropy.

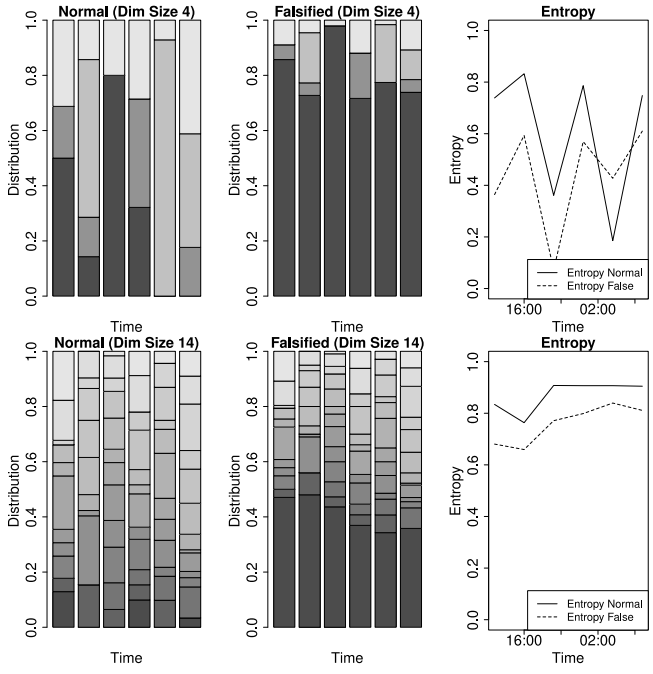


Fig. 9. Distribution and entropy for vectors of size of 4 and 14.

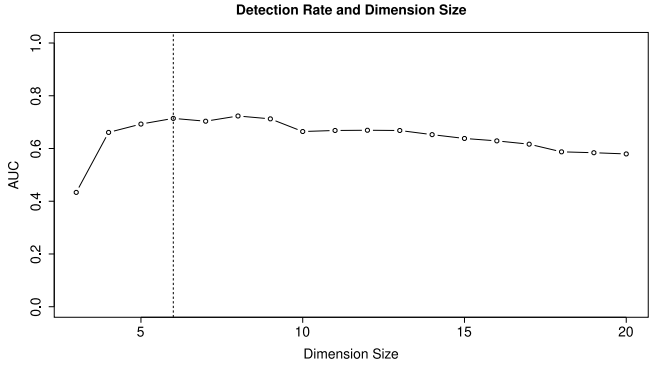


Fig. 10. AUC for dimension sizes.

the entropy of anomalous and normal values. We evaluate this using the Area under the Curve of a ROC Curve. A ROC Curve shows the sensitivity (true positive rate) and the specificity (true negative rate) for any possible threshold to divide benign and anomalous entropy values. The accuracy of a method can be defined by the area under the curve (AUC). A good method can maximize both sensitivity and specificity, which results in a big area under the curve, while a random method results in a diagonal line and an AUC of 0.5. In Fig. 10, we can see the area under the curve for different dimension sizes (each showing the AUC for 240 classifications = 40 days), which shows no significant improvement with dimension sizes greater six (vertical dotted line). *Note that, the AUC still improves with more dimensions if we include the prediction. We still used a dimension size of six due to practical reasons.*

4.4. Parameter evaluation: Prediction

In the previous section, we evaluated the influence of feature window size and dimension sizes on the expected difference of the entropy for normal and anomalous energy demand. We found that, even in case of good conditions the difference of benign and tampered data is often smaller than the expected standard deviation of the entropy, which is due to the regular patterns. Hence, in order to detect anomalies more reliably, we need to remove the usual time-depending patterns carried by our features and the resulting entropy. In the following, we use a time series prediction algorithm to forecast the expected entropy. By subtracting the predicted from the actual value we aim to receive a straight line, without time-depended pattern, which is high for normal and low for falsified data.

For our experiments, we used Holt–Winters, which models the level, seasonality and slope of a time series using training data. A small amount of training data is preferable as privacy concerns and performance may not allow huge sets of historic data for ex post analysis. However, small amounts of training data can lead to over-fitting: while the Holt–Winters model perfectly fits the training data, the actual data would show unpredicted patterns different from the learned data and not match the model.

The length of training data does depend upon previous parameters, such as the window size and dimension, as those parameter define the frequency of regular pattern. Here, we chose the root mean square error (RMSE) between model and actual data as a

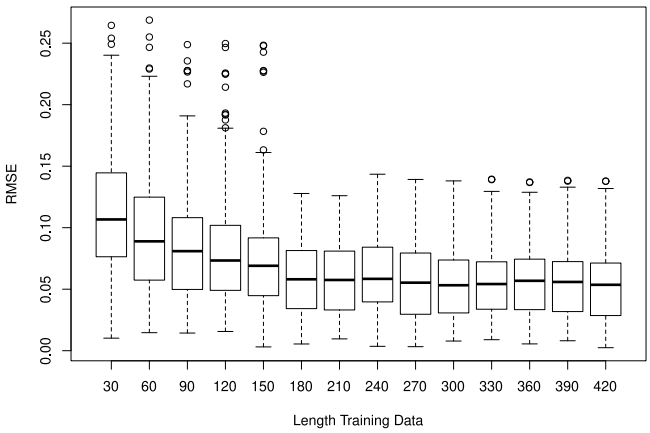


Fig. 11. RMSE for different lengths of training data.

Table 1
RMSE for each feature and dimension.

H_1	DM	p-value
$f_1 < f_2$	7.7143	4.012e−13
$f_1 < f_3$	7.635	6.393e−13
$f_3 < f_2$	3.6171	0.0001934

rough estimation of the model quality. The RMSE, which is the square root of the variance of the residuals, shows how close the model fits to the actual data. It is in the same unit as the training data, the normalized entropy with a range [0, 1], which means we can expect an RMSE of [-1,1]. Values close to zero indicate better fit, as zero means that we do not have any variance between training data and model.

Fig. 11 shows boxplots for the RMSE, ranging from 30 entropy values (5 days) to 420 values (70 days) as training data (avg. of 300 trials). We used the feature 'High Amplitude' on dimension 'Date' with a window size of 48 power measurements (4 h) and a dimension size of 6 to compute the entropy. In this specific case, the RMSE is minimized at 180 entropy values, which is about a month.

In Table 1, we aim to compare our introduced features to find out which feature is in general better to predict. For this purpose, we set up a Diebold–Mariano test, which is using the forecast error to compare the accuracy of two forecasts with a so called DM statistic. The DM statistic is used to compute a hypothesis test with the null hypothesis that both forecasts are of equal accuracy, and the alternative hypothesis that one method has greater accuracy. A small p-value indicates strong evidence against the null hypothesis, which means that the null hypothesis can be rejected and the alternative hypothesis is true. A large p-value indicates weak evidence against the null hypothesis, which means that no conclusion can be drawn. Table 1 shows the resulting hypothesis test for the number of high amplitude points (f_1), the number of amplitude changes (f_2) and the number of similar amplitudes in a row (f_3), with the same parameters that were used in Fig. 11.

5. Detection rate and performance

We claim that the advantage of our entropy metric is especially the ability to keep information on outliers after aggregating several values, whereas the aggregate value can be utilized to apply time series prediction. For this reason, we first compare our metric with an alternative metric and evaluate the value of the time series prediction. In the second part of this section, we compare our complete anomaly detection scheme with other state of the art approaches.

5.1. Performance of the metric

Here, we evaluate if the time series prediction increases our detection rate in contrast to the same method without prediction. Furthermore, we check if the entropy is really a good method to aggregate data while preserving outlier. As an alternative to the entropy, we compute the maximum distance to the mean-value of a row: in contrast to the entropy, this value is large in case of big outliers and small if all values are uniform distributed – we call this method *D2M*. Given is, for both methods (D2M and Entropy), a time series consisting of individual values (our metric), which summarize several dimensions over a predefined time window. Each method is evaluated in combination with Holt–Winters (a) and without (b). The decision algorithm for both methods differs slightly: for (a), a value is anomalous if the metric is exceeds a threshold. For (b), a value is anomalous if the actual metric minus prediction exceeds a threshold.

We evaluate the AUC of both methods with a so called bootstrap test. Here, the AUC is repeatedly ($N = 2000$) computed with the original inputs for the ROC curve re-sampled, which approximately follows a normal distribution used to perform a hypothesis test. The null hypothesis is that the true difference between both AUC is zero and the alternative hypothesis is that method one performs better than method two. A small p-value shows that the null hypothesis can be rejected. Table 2, at the end of the paper, shows the result of this experiment for each feature, the number of high amplitude points (f_1), the number of amplitude changes (f_2) and the number of similar amplitudes in a row (f_3), with dimension size 6 and a window size of 48. Furthermore, the table shows the AUC for one, two or three corrupted samples (days). *Note that, a majority of falsified samples cannot be detected per definition.* To train Holt–Winters, we predicted 6 entropy values (1 day) from the last 60 entropy values (2 month). Each experiment was repeated (over a time of 4 month) to get 120 prediction values, which are used to construct the AUC of a ROC Curve. For simplicity, we interpret the AUC as accuracy, where 0.5 is random and 1 is perfect. In the results we can see, that the time series prediction significantly improves the results for both metrics, the entropy and D2M. Although our proposed method, the entropy-inspired metric in combination with time series prediction, results in acceptable detection rates mostly over 90%, which can compete with the other function, it only performs better if we have several outliers in the distribution – which means that e.g. the current day and day before show energy theft. A case with several outliers is difficult to detect for alternative function.

The lower part of the table shows the same results for dimension 'House'. Unfortunately, due to the different load pattern in the dimension 'House', two of our three proposed features were not working well enough for a practical usage. For features 'High Amplitude' and 'Similar in a Row', we were not able to optimize the parameter ε well enough to get an approximately uniform or otherwise predictable pattern each household. However, the feature 'Amplitude Changes' (f_2), was resilient to these different load patterns and performed very well (This is not surprising, because the experiments in previous sections already pointed out that the $\hat{corr}$ and RMSE perform best for feature 'Amplitude Changes').

The results for type 2 falsified data in Table 3 are similar. We expected that some features react better to tampering on the amplitude and others on changed load patterns, but that was not the case. All features react very well on changed load patterns (type 1) and worse if the amplitude not affects load patterns (type 2). 'Amplitude Changes' performed well, while both other features did not detect energy theft in this case and can only work if the load patterns of the data sources are very similar.

Table 2
AUC: type 1 falsified.

		Dimension: Date						
		f_1	f_2	f_3		f_1	f_2	f_3
1 of 6	Ent+TS	0.953	0.951	0.938	Ent	0.76	0.684	0.661
	D2M+TS	1	0.998	0.938	D2M	0.973	0.854	0.761
	p-value	0.014	0.00607	0.99	p-value	4.1e−09	5.98e−10	0.000323
2 of 6	Ent+TS	0.942	0.981	0.907	Ent	0.767	0.771	0.624
	D2M+TS	0.893	0.921	0.742	D2M	0.626	0.675	0.523
	p-value	0.0547	0.0229	5.4e−06	p-value	2.8e−05	7.37e−08	0.333
3 of 6	Ent+TS	0.853	0.975	0.822	Ent	0.646	0.756	0.557
	D2M+TS	0.621	0.853	0.518	D2M	0.657	0.587	0.672
	p-value	1.31e−07	0.00308	3.6e−11	p-value	0.911	1.13e−12	0.251
		Dimension: House						
		f_1	f_2	f_3		f_1	f_2	f_3
1 of 6	Ent+TS	0.456	0.87	0.591	Ent	0.68	0.83	0.507
	D2M+TS	0.689	0.926	0.669	D2M	0.461	0.94	0.633
	p-value	0.0173	0.00105	0.103	p-value	1.74e−08	0.000145	0.206
2 of 6	Ent+TS	0.561	0.956	0.609	Ent	0.729	0.897	0.707
	D2M+TS	0.629	0.956	0.531	D2M	0.818	0.894	0.577
	p-value	0.127	0.944	0.0876	p-value	0.00223	0.894	0.00163
3 of 6	Ent+TS	0.672	0.876	0.682	Ent	0.851	0.815	0.76
	D2M+TS	0.758	0.829	0.589	D2M	0.927	0.731	0.644
	p-value	0.0368	0.0679	0.0137	p-value	0.0029	0.00103	0.00129

Table 3
AUC: type 2 falsified.

		Dimension: Date						
		f_1	f_2	f_3		f_1	f_2	f_3
1 of 6	Ent+TS	0.912	0.792	0.819	Ent	0.73	0.508	0.497
	D2M+TS	0.998	0.852	0.838	D2M	0.969	0.654	0.556
	p-value	0.000932	0.0713	0.637	p-value	6.47e−10	2.04e−05	0.0272
2 of 6	Ent+TS	0.939	0.879	0.673	Ent	0.749	0.378	0.449
	D2M+TS	0.936	0.846	0.584	D2M	0.645	0.404	0.601
	p-value	0.857	0.386	0.0128	p-value	0.00204	0.139	0.14
3 of 6	Ent+TS	0.847	0.843	0.65	Ent	0.644	0.607	0.4
	D2M+TS	0.686	0.797	0.545	D2M	0.629	0.501	0.677
	p-value	8.4e−05	0.102	0.294	p-value	0.874	0.319	0.00517
		Dimension: House						
		f_1	f_2	f_3		f_1	f_2	f_3
1 of 6	Ent+TS	0.456	0.87	0.591	Ent	0.68	0.83	0.507
	D2M+TS	0.689	0.926	0.669	D2M	0.461	0.94	0.633
	p-value	0.0173	0.00122	0.102	p-value	1.47e−08	7.89e−05	0.206
2 of 6	Ent+TS	0.561	0.956	0.609	Ent	0.729	0.897	0.707
	D2M+TS	0.629	0.956	0.531	D2M	0.818	0.894	0.577
	p-value	0.126	0.944	0.0701	p-value	0.00165	0.895	0.00127
3 of 6	Ent+TS	0.672	0.876	0.682	Ent	0.851	0.815	0.76
	D2M+TS	0.758	0.829	0.589	D2M	0.927	0.731	0.644
	p-value	0.0327	0.0755	0.0153	p-value	0.00274	0.000817	0.0016

We believe that the performance may be further increased by clustering similar load curves (in order to reduce the complexity of daily patterns) – which is not in the scope of this work. The feature 'Amplitude Changes' showed results without clustering households according to their similarity and performed well solely because the pattern of each individual house was consistent enough. A generalization to other data sets is difficult, but we showed that our concept works on the condition that we find a feature with consistent pattern on each data source, which can generate outliers in case of manipulated data, and showed methods for parameter optimization to ensure that these conditions are adhered.

5.2. Performance of the detection scheme

Here we introduce a final comparison of our scheme with two other state of the art anomaly detection methods on energy demand, namely a method inspired by AMIDS from Mclaughlin et al. [47], which models the energy consumption behavior of a household using Naive Bayes, and a method based on XMR charts from Spiric et al. [26]. Mclaughlin's original article utilizes so-called Non Intrusive Appliance Load Monitoring (NIALM) profiles to associate each on/off amplitude in a households energy load curve to a certain appliance. The three resulting vectors with amplitudes, appliance names and on/off operations are used as input for the supervised learning of the Naive Bayes algorithm, which computes the probability for energy theft for each data point. In contrast to our previous method, AMIDS has a strict requirement for high data resolution. If the amplitudes of individual devices are not visible, the detection rate is highly decreased. For our experiment, we had to simplify Mclaughlin's method, because the setup of a suitable NIALM database is beyond our scope. Note that, Mclaughlin evaluated his method with an energy demand simulation, where the mapping of appliances to predefined profiles is generally easier. Here, we use a simple clustering algorithm

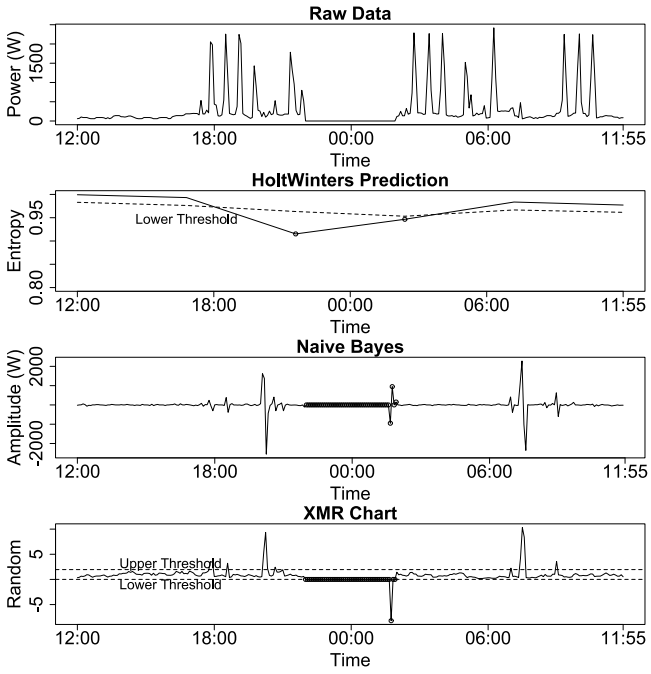


Fig. 12. Overview of anomaly detection methods.

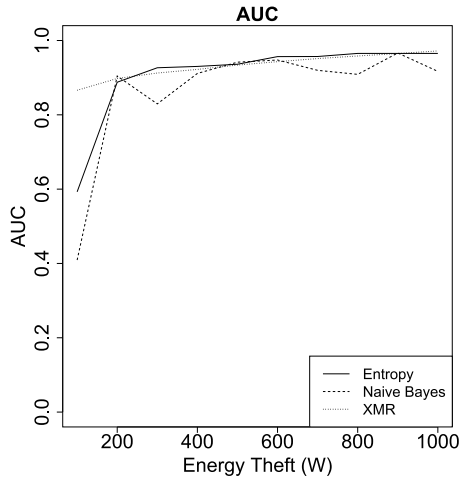


Fig. 13. AUC depending on the amount of energy theft.

Table 4
AUC: type 1 falsified.

	AUC		AUC		AUC
EntTS	0.965	EntTS	0.965	Naive	0.962
Naive	0.962	XMR	0.972	XMR	0.972
p-value	0.859	p-value	0.712	p-value	0.475

instead of an elaborated, hand-labeled NIALM database. Hence, we assume that the most frequent amplitudes (as arranged by the cluster centers) correspond to the amplitudes of different devices. One of the limitations of such a method is that appliances with similar amplitude and appliances used together cannot be recognized as individual device. Our experiments indicate that the results are still consistent enough for anomaly detection since we can get a detection rate similar to Mclaughlin's. However, it may be possible to further increase the detection rate of this method by using a better NIALM algorithm and higher resolution data. Spiric's fraud detection is based on monitoring the 'random component' of the energy demand, which means that the raw input data is decomposed into a seasonal component, trend component and random component (e.g. by using a moving average time series decomposition algorithm). In order to define the threshold for energy theft, Spiric utilizes a so called XMR chart, which computes an upper and lower limit using the mean moving range. Note that, the threshold of the XMR chart is not relevant for our results, because we use the AUC as metric (see Section 5.1). The AUC computes the result for any possible threshold, which means our results of Spiric's method may be slightly better than with a fixed threshold.

Fig. 12 shows an overview of these methods: The top plot shows the raw input data with four hours of energy theft around 0:00 pm. Next we see our own method, which computes the entropy (black line) and the Holt Winters prediction of the expected entropy (dashed line) as a lower threshold. Here, energy theft results in a small entropy. The third plot shows Mclaughlin's method, which is using a time series of amplitudes as input data and assigns a probability for energy theft, which is the output of Naive Bayes, to each measurement. Here, energy theft results in a high probability. The bottom plot shows the XMR chart with the random component of time-series decomposition as input, which detects energy theft with a lower threshold (dashed line). Here, energy theft results in a small or negative number relative to the mean of the random component.

Note that, in order to compare these different methods, we had to accept some limitations. E.g. it may be possible that, especially the Naive Bayes method, can be improved with higher resolution data, because the edges of on/off operations are more visible. Furthermore, it may be possible to optimize the lengths of input data, e.g. the amount of training data or length of the expected seasonality for the decomposition algorithm. For the Naive Bayes we used the previous day as training data, because using more training data worsened the results. For the decomposition, we used the default setting of one day seasonality.

Since our own method has a low resolution output data, we had to aggregate the output of both other methods to the same resolution in order to compute a consistent AUC. Both methods utilize only a single source of input data, and hence it is not possible to check the results of multiple compromised sources, which is one of the strong points of our own algorithm. Since the previous section already contains a detailed evaluation of our own method (see Table 2), we only conducted this experiment for the feature 'Amplitude Changes'.

Fig. 13 shows the results for different amounts of energy theft. Each algorithm aimed to detect one day of energy theft on ECO data household 1 in August. The plot shows the mean of 15 experiments. Note that, the output of Naive Bayes still has a lot of variance because the result depends on the random cluster centers used to determine the appliances. The x-axis shows the AUC and the y-axis shows the amount of power subtracted from the original energy demand. We can see here, that Spiric's method is especially good at detecting smaller amounts of energy theft.

In Table 4, we perform the same bootstrap test as in section Section 5.1 with the alternative hypothesis that the accuracy of both methods is different. However, for the type 1 energy theft, we cannot clearly reject the null hypothesis that all methods perform equally good, as all methods have a high accuracy. As our method is generally intended to use additional data sources which are not considered in the other two methods, we believe that these three methods should be used together to complement each other.

6. Conclusion

In this work, we showcased anomaly detection in different dimensions, which can unveil, otherwise hidden, anomalous data

if the majority of data in a single source is compromised. The prerequisite to utilize multiple dimensions is an outlier producing feature comparable with other data sources as well as an aggregation method, which preserves these outliers, to remove repeating patterns. We showed three such features and a systematic approach to fine-tune and adjust them, which significantly affects the detection rate. We demonstrated that high level information, such as the 'activity', can be utilized to normalize data to a fixed range and fine-tune the feature to the specific 'normal' characteristics of a data source. For the fine-tuning, we illustrated how to find an optimal threshold for each feature, to distinguish normal and falsified data. Next, we showed how to find the optimal window size for each feature, to minimize the standard deviation of normal data. Furthermore, we evaluated the influence of the number of data sources, so that, after the aggregation, falsified data still produces outliers which are greater than the standard deviation of normal data. In order to remove the repeating pattern from the aggregated data, we demonstrated how to find the optimal length of training data to maximize the predictability of the metric. Our approach of extensive parameter tuning, to adapt the feature to the specifics of a data source and a certain malicious activity, may be seen as limitation. However, we prioritized examination of the statistical influence of parameters over automation, because we argue that anomaly detection can only work with a solid understanding of the underlying data. A limitation specific to our anomaly detection method is the low output resolution, which was required to reduce the standard deviation of the feature. Colloquially speaking, it means that the majority of the electricity in this time window must be manipulated for a detection.

For our two scenarios of energy theft, we had detection rates above 90%, whereas the number of amplitude changes $\geq \varepsilon$ performed especially well. We were able to detected tampered data even by utilizing different households, which were not clustered according to their similarity, as data source. Removing daily patterns with Holt–Winters significantly improved the detection rate from about 75% to above 90%. Apart from the entropy-inspired metric, other aggregate methods may work as well, but our entropy-inspired metric is especially robust in presence of multiple outliers. The detection rate of the alternative aggregation method (D2M) decreased up to 10% for each additional compromised data source, while the detection rate of the entropy-inspired metric did not significantly drop with up to half of all data sources compromised. Sophisticated and stealthy tampering methods, were not analyzed in this work, as the detection of data mimicking attacks is a well known challenge and inherent limitation of anomaly detection and beyond our scope. As future prospects, we suppose to evaluate the usage of additional measurements, which are correlated to the power, to complicate the construction of such legitimate looking false data. Furthermore, it would be interesting to discuss the selection approach of data sources under different conditions, e.g. depending on the similarity of data sources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the German Federal Ministry for Economic Affairs and Energy under the Central Innovation Programme for SMEs (ZIM) Grant No. ZF4131804HB8.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.segan.2019.100290.

References

[1] A. Illera, J. Vidal, Lights Off! the Darkness of the Smart Meters, BlackHat Europe, 2014.
[2] D.R. Bohi, M.B. Zimmerman, An update on econometric studies of energy demand behavior, Annu. Rev. Energy 9 (1) (1984) 105–154.
[3] W. Kleiminger, C. Beckel, S. Santini, Household occupancy monitoring using electricity meters, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2015, pp. 975–986.
[4] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, D. Irwin, Private memoirs of a smart meter, in: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, ACM, 2010, pp. 61–66.
[5] Y.G. Yohanis, J.D. Mondol, A. Wright, B. Norton, Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use, Energy Build. 40 (6) (2008) 1053–1059.
[6] P. Price, Methods for Analyzing Electric Load Shape and its Variability, Lawrence Berkeley National Laboratory, 2010.
[7] A. Druckman, T. Jackson, Household energy consumption in the uk: A highly geographically and socio-economically disaggregated model, Energy Policy 36 (8) (2008) 3177–3192.
[8] J. Carroll, S. Lyons, E. Denny, Reducing household electricity demand through smart metering: The role of improved information about energy saving, Energy Econ. 45 (2014) 234–243.
[9] F. Mcloughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study, Energy Build. 48 (2011) http://dx.doi.org/10.1016/j.enbuild.2012.01.037.
[10] J. Kolter, J. Ferreira, A large-scale study on predicting and contextualizing building energy usage, in: Proceedings of the National Conference on Artificial Intelligence, vol. 2, 2011, pp. 1–8.
[11] C. Beckel, L. Sadamori, S. Santini, Automatic socio-economic classification of households using electricity consumption data, in: Proceedings of the Fourth International Conference on Future Energy Systems, ACM, 2013, pp. 75–86.
[12] A. Newing, B. Anderson, A. Bahaj, P. James, The role of digital trace data in supporting the collection of population statistics–the case for smart metered electricity consumption data, Popul. Space Place 22 (8) (2016) 849–863.
[13] P. Jain, P. Tripathi, Scada security: a review and enhancement for dnp3 based systems, CSI Trans. ICT 1 (4) (2013) 301–308.
[14] A.R. Metke, R.L. Ekl, Smart grid security technology, in: Innovative Smart Grid Technologies, Vol. 2010, ISGT, IEEE, 2010, pp. 1–7.
[15] W. Westerhof, Horus scenario - exploiting a weak spot in the power grid, 2017, URL https://horusscenario.com/.
[16] A. Dabrowski, J. Ullrich, E.R. Weippl, Grid shock: Coordinated load-changing attacks on power grids: The non-smart power grid is vulnerable to cyber attacks as well, in: Proceedings of the 33rd Annual Computer Security Applications Conference, ACM, 2017, pp. 303–314.
[17] W. Wang, Z. Lu, Cyber security in the smart grid: Survey and challenges, Comput. Netw. 57 (5) (2013) 1344–1371.
[18] V. Delgado-Gomes, J.F. Martins, C. Lima, P.N. Borza, Smart grid security issues, in: Compatibility and Power Electronics (CPE), 2015 9th International Conference on, IEEE, 2015, pp. 534–538.
[19] J. Anu, R. Agrawal, C. Seay, S. Bhattacharya, Smart grid security risks, in: Information Technology-New Generations (ITNG), 2015 12th International Conference on, IEEE, 2015, pp. 485–489.
[20] A. Nizar, Z. Dong, Y. Wang, Power utility nontechnical loss analysis with extreme learning machine method, IEEE Trans. Power Syst. 23 (3) (2008) 946–955.
[21] J. Nagi, K. Yap, S. Tiong, S. Ahmed, A. Mohammad, Detection of abnormalities and electricity theft using genetic support vector machines, in: TENCON 2008-2008 IEEE Region 10 Conference, IEEE, 2008, pp. 1–6.
[22] S.S.S.R. Depuru, L. Wang, V. Devabhaktuni, Support vector machine based data classification for detection of electricity theft, in: 2011 IEEE/PES Power Systems Conference and Exposition, IEEE, 2011, pp. 1–8.
[23] A.A. Cárdenas, S. Amin, G. Schwartz, R. Dong, S. Sastry, A game theory model for electricity theft detection and privacy-aware control in ami systems, in: 2012 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton, IEEE, 2012, pp. 1830–1837.
[24] C. Bandim, J. Alves, A. Pinto, F. Souza, M. Loureiro, C. Magalhaes, F. Galvez-Dur, Identification of energy theft and tampered meters using a central observer meter: a mathematical approach, in: 2003 IEEE PES Transmission and Distribution Conference and Exposition (IEEE Cat. No. 03CH37495), Vol. 1, IEEE, 2003, pp. 163–168.

[25] S. Salinas, M. Li, P. Li, Privacy-preserving energy theft detection in smart grids: A p2p computing approach, IEEE J. Sel. Areas Commun. 31 (9) (2013) 257–267.
[26] J.V. Spirić, M.B. Dočić, S.S. Stanković, Fraud detection in registered electricity time series, Int. J. Electr. Power Energy Syst. 71 (2015) 42–50.
[27] D.I. Urbina, J.A. Giraldo, A.A. Cardenas, N.O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, H. Sandberg, Limiting the impact of stealthy attacks on industrial control systems, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016, pp. 1092–1105.
[28] J. Bouché, D. Hock, M. Kappes, On the performance of anomaly detection systems uncovering traffic mimicking covert channels, in: INC, 2016, pp. 19–24.
[29] D.E. Denning, An intrusion-detection model, Softw. Eng. IEEE Trans. 13 (2) (1987) 222–232.
[30] Y. Zhang, N. Meratnia, P. Havinga, Outlier detection techniques for wireless sensor networks: A survey, Commun. Surv. Tutor. 12 (2) (2010) 159–170.
[31] M. Xie, S. Han, B. Tian, S. Parvin, Anomaly detection in wireless sensor networks: A survey, J. Netw. Comput. Appl. 34 (4) (2011) 1302–1325.
[32] H. Braun, S.T. Buddha, V. Krishnan, A. Spanias, C. Tepedelenlioglu, T. Yeider, T. Takehara, Signal processing for fault detection in photovoltaic arrays, in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, 2012, pp. 1681–1684.
[33] S. Dienst, J. Schmidt, S. Kühne, Case study: Condition assessment of a photovoltaic power plant using change-point analysis, in: SMARTGREENS, 2013, pp. 159–164.
[34] S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, S. Zonouz, A multi-sensor energy theft detection framework for advanced metering infrastructures, IEEE J. Sel. Areas Commun. 31 (7) (2013) 1319–1330.
[35] M. Raciti, S. Nadjm-Tehrani, Embedded cyber-physical anomaly detection in smart meters, in: Critical Information Infrastructures Security, Springer, 2013, pp. 34–45.
[36] B. Rossi, S. Chren, B. Buhnova, T. Pitner, Anomaly detection in smart grid data: An experience report, in: 2016 Ieee International Conference on Systems, Man, and Cybernetics, smc, IEEE, 2016, pp. 002313–002318.
[37] L. Mookiah, C. Dean, W. Eberle, Graph-based anomaly detection on smart grid data, in: The Thirtieth International Flairs Conference, 2017, pp. 1–6.
[38] S.C. Yip, W.N. Tan, C. Tan, M.T. Gan, K. Wong, An anomaly detection framework for identifying energy theft and defective meters in smart grids, Int. J. Electr. Power Energy Syst. 101 (2018) 189–203.
[39] Z. Fengming, L. Shufang, G. Zhimin, W. Bo, T. Shiming, P. Mingming, Anomaly detection in smart grid based on encoder–decoder framework with recurrent neural network, J. China Univ. Posts Telecommun. 24 (6) (2017) 67–73.
[40] T. Andrysiak, Ł. Saganowski, P. Kiedrowski, Anomaly detection in smart metering infrastructure with the use of time series analysis, J. Sens. 2017 (2017).
[41] Y. Zhou, H. Zou, R. Arghandeh, W. Gu, C.J. Spanos, Non-parametric outliers detection in multiple time series a case study: Power grid data analysis, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 1–8.
[42] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Heart Circ. Physiol. 278 (6) (2000) H2039–H2049.
[43] I. Vranken, J. Baudry, M. Aubinet, M. Visser, J. Bogaert, A review on the use of entropy in landscape ecology: heterogeneity, unpredictability, scale dependence and their links with thermodynamics, Landsc. Ecol. 30 (1) (2015) 51–65.
[44] A. Wagner, B. Plattner, Entropy based worm and anomaly detection in fast ip networks, in: 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, WETICE'05, IEEE, 2005, pp. 172–177.
[45] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[46] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, S. Santini, The eco data set and the performance of non-intrusive load monitoring algorithms, in: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, ACM, 2014, pp. 80–89.
[47] S. McLaughlin, B. Holbert, S. Zonouz, R. Berthier, Amids: A multi-sensor energy theft detection framework for advanced metering infrastructures, in: 2012 IEEE Third International Conference on Smart Grid Communications, SmartGridComm, IEEE, 2012, pp. 354–359.