M.Sc Thesis: Detecting Anomalies, Attacks and intrusions in smart grids

Name: Arun Kumar Naranahalli Anjanappa
Matrikelnummer: 2372024
Fachbereich: Informatik (Distributed Software Systems)
Betreuer: Prof. Dr. Max Muhlhauser
Zweitbetreuer: M Sc. Carlos Garcia C

Introduction

For many years there has been no change in the basic structure of the power grids. Meeting the rising demand for electricity, measures to reduce equipment failures, reduction in the available fossil fuels, energy storage problems and the one way communication have posed a serious challenges on the electricity and transportation industries as described by [18]. The energy grid has evolved from a centralized infrastructure to a complex system where every level of the distribution pipeline comprises of multiple smart devices which can connect to multiple power sources that can produce energy, as well as store it and exchange it with many other smart devices. The centralized nature of the power grid has been exchanged with the distributed systems which not only distribute the power but also exchange information which can improve efficiency, reliability and safety of the power grids. The existing grid is lack of communication capabilities, while a smart power grid infrastructure is full of enhanced sensing and advanced communication and computing abilities as illustrated in Figure 1.

Power grids evolved with the integration of intelligent infrastructures and communication technology. These technologies not only made the power grids smart but also helped to achieve the goal of meeting the high demand power with the connection of energy produced from renewable resources and reducing the loss by employing smart technologies for detecting and minimizing power outages. These grids combine the power and communication networks to connect homes, offices, and factories (consumers) to multiple distributed power providers (small-scale power generators) such as solar, wind, fuel cells, and facilities that store generated power. Smart grids will provide more electricity to meet rising demand, increase reliability and quality of power supplies, increase energy efficiency, integrate low carbon energy sources into power networks. Smart grids possess demand response capacity to help balance electrical consumption with supply, as well as the potential to integrate new technologies to enable energy storage devices and the large-scale use of electric vehicles [1]. Until the rise of Wi-Fi and mobile technology, power systems had very limited resources to detect the problem areas and resolve. Today, these capabilities are increasingly being employed by the power grids to give you update of the outages as soon as possible. The integration of smart devices to existing power distribution lines helps to monitor the state of the power grid, the operational conditions of the communication lines and respond for outages. Information are being communicated continuously determining the current operational conditions of the distribution lines, adding to the reliable transfer of power to the customers. Intelligent sensors deployed in smart grid network reduces the number of customers being impacted with the help of real-time outage response, which in turn restores the power to the affected areas in an alternative path. Addition to the automatic restoration, the affected area is isolated from the distribution network quickly which may have caused due to multiple reasons such as severe weather conditions, natural calamities and link breakages.

Data generated from smart grids: Delete this line after completing

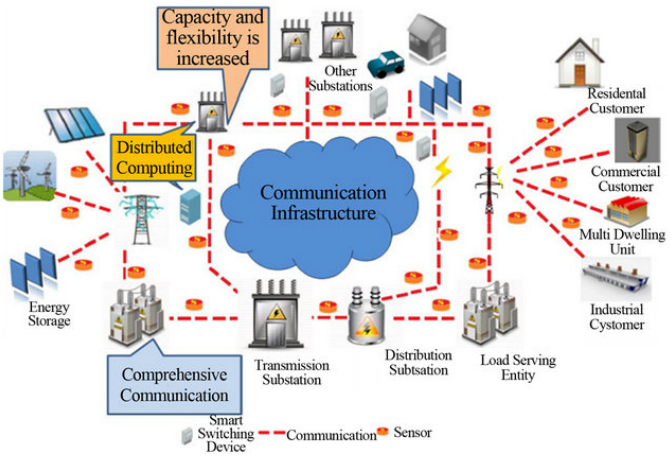

Figure 1: Smart grid architecture increases the capacity and flexibility of the network and provides advanced sensing and control through modern communications technologies [18]

Apart from increased efficiency, self-restoration, operation automation, and renewable energy integration, the smart grids are potential sources of big data. The information technology support for the power transmission and distribution produces rapid amount of data. The sensors, the smart meters, the control units and the monitoring devices all add up to this huge volumes. The grid is becoming more and more digital with the information recorded by the sensors and the smart meters deployed at the users' premises and along the grid and power stations. This data can be for example, consumption data from smart metering and electric vehicles, data telling about the conditions of the components in the grid [3]. For achieving fine-grained monitoring and scheduling, information from the power grid needs to be collected within short intervals. Data are generated from various activities such as: power utilization habits of the users; phasor measurement data for situational awareness; energy consumption data measured by the widespread smart meters; energy market pricing and bidding data collected by advanced metering infrastructure; management, control and maintenance data of devices and equipment in the power generation, control data exchanged between transmission and distribution networks acquired by intelligent electronic devices [24]. **Energy saving using Smart grids**

Effects of data, why are they collected : Delete this line after completing

Processing and analyzing these data reveals deeper insights that can help expert to improve the operation of power grid to achieve better performance. The technology to collect massive amounts of data is available today, but managing the data efficiently and extracting the most useful information out of it remains a challenge. Data obtained from the smart grids provides valuable information which helps for efficient grid operations. It is also closely related to the safety

and reliability of the power system operation and management based on data-driven decision support. Using data from meters and sensors, an operator could be alerted that a transformer is having problems and help to make a decision from the information provided. For example, in the past when a transformer would fail, the lights would go out and the operator only knew about it when customers called. In future we would wish to have a sensor on the power line which sends an alert to the operator about that transformer. The operator recognizes where the outage is and dispatches a person to fix the problem. This gets the lights back on sooner and results in more satisfied customers [40]. With this technology, utilities are equipped to deliver power more efficiently, improve operations, reduce emissions and management costs, and restore power faster. And operators are able to immediately identify outages, allowing for improved efficiency to manage responses.

Smart grid data could also generate information on how individual customers respond to requests of consumption reduction. It is also possible to use real-time metering data to discover unaccounted consumption when energy is being diverted and stolen, reducing the cost of distribution operations. With smart meters, utilities can learn about the information of the number of units consumed by the user and the state of the grid. This information allows the power grids to deploy optimization techniques to find the optimal power generation, control mechanisms and transmission strategies for that state of grid. Construction and application of the smart grid have stimulated the accumulation of electric power grid operation data, production management data and electricity consumption data [11]. These accumulated data contain redundant, missing and outlier data, resulting in serious issues of electric power data quality. These electrical data can be used by data analytics researchers to obtain patterns and to make better decisions for electric utilization and control. They can make use of several techniques, including statistics, clustering algorithms, data mining, machine learning, signal processing, pattern recognition, optimisation and visualisation method to capture, curate, analyze and visualize the data.

As there are advancement in technology, there are adverse effects as well. Over the last years power grids have been targeted with hazardous malware to steal their data and get access to their communication. Security is one feature which plays a very huge role in every communication entity. Since smart grids and the power generators exchange informations, intrusions, attacks and anomalies could occur because of malware activities. In order to accomplish these security goals we propose to use techniques which will be discussed in the later sections.

Our proposal for nullifying one of the effect: Delete this line after completing The massive amounts of measurement data that will be made available at distributed locations that can and must be leveraged to optimally operate the power grid.

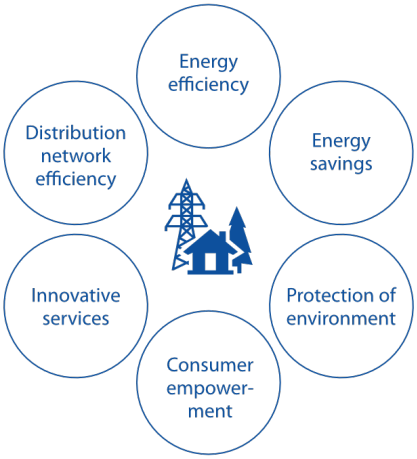

Figure 2: Six ways the adoption of smart metering systems can benefit the electricity customer. Image src: http://www.foxconngfo.com/iot/foxconn/page-11.html

Motivation

Added Newly: Should i write about how electricity is produced and transmitted, what are the tradeoffs to be considered when deploying the anomaly detection algorithms in the smart grids.

In order to efficiently utilize the energy it requires one to study detailed knowledge of the energy usage patterns of the locations. For this we consider the energy usage patterns of the locations in order to identify irregularities of the transmitted energy. We do this by detection algorithms using the modern machine learning techniques. Apart from the economic benefit brought along by the increased energy efficiency, there is also an urgent global need to reduce carbon emissions such as carbon dioxide, which is closely tied up with the generation and hence consumption of electricity [9]. Generally, this can be reduced by using clean (or cleaner) energy sources and the control of electricity consumption or demands at the users side. The main aim of smart grids are to provide

- more efficient transmission of electricity.
- quicker restoration of electricity after power disturbances.
- reduced operations and management costs for utilities, and ultimately lower power costs for consumers.
- reduced peak demand, which will also help lower electricity rates.
- increased integration of large-scale renewable energy systems.

- better integration of customer-owner power generation systems, including renewable energy systems.
- improved security.

Nevertheless, increased interconnection and integration also introduce vulnerabilities into the grid. Failure to address these problems will hinder the modernization of the existing power system. Through advanced sensing technologies and control methods, smart grids can capture and analyze data regarding power usage, delivery, and generation in near real-time. According to the analysis results, the smart grid may provide predictive information and corresponding recommendations to all stakeholders (e.g., utilities, suppliers, and consumers) regarding the optimization of their power utilization. It may also offer services like intelligent appliance control for energy efficiency and better integration of distributed energy resources (DERs) to reduce carbon emissions. Apparently, it is not a simple grid in the sense of our current power grid. It can be regarded as a system of systems that involves both information technology (IT) and electricity system operations and governance. Such a complex system undoubtedly presents many challenges, especially in cyber security and privacy aspects [27].

Smart grids should also consider the wide range of needs and power quality requirements. Power quality involves factors like voltage flicker, voltage volume, momentary interruptions, etc. Different consumers may have distinct power quality requirements (e.g., industrial vs. residential users). Optimizing the distribution of power by detecting the irregularities, outliers or anomalies will reduce power consumption. These timely detection of faults by using the proposed technologies will help in operating resiliently to disturbances, attacks and Natural disasters. Gann et al. affirmed that cities become smarter when they exploit the increasing data and analytical techniques available, in order to improve energy effectiveness and efficiency through the integration of physical infrastructures and digital technologies [16]

end

The energy grid can now interact with the end user to control his energy consumption, by either direct control on some of his appliances (for example, the washing machine) or indirect control, by providing fine grain resolution on the cost of the energy at a given day and time, such that the final user tunes up his own schedule for the energy consumption. The grid can also promote cooperation among different prosumers (both producers and consumers of energy) to enable more efficient energy usage, particularly in what concerns consuming on the spot for renewable energies.

Discovering new information can be very beneficial to both power distribution grids and customers. Uncovering hidden energy usage patterns, identifying untapped energy efficiency opportunities, engaging customers for demand response programs by discovering high usage patterns, for instance, inefficient thermostat setpoints and charging their electric vehicles, can save a lot of energy and expenses of the customers [32].

Data is the fundamental currency of the smart grid. A clear understanding of how these data are generated, what it consists of, and the benefits it can be

used to deliver is critical to realizing the fullest possible returns from smart grid investments. The benefit is to have enriched information of the performance of the system, its stability, and customer consumption. For this task, anomalies are of special interest, because they can be caused either by faulty equipment or potentially misconfigured devices consuming significantly more or less energy than required for proper operation.

Complex electric power system contains large amounts of real-time data. Data is accurate or not, which determines the safety and reliability of electric power system. Outlier data may affect the normal operation of electric power system and even threaten the security of entire system. Hence, in order to ensure the stable and safe operation of electric power system, it has important significances to extract and detect these outlier data from the original data. The significances of electricity consumption outlier data have two aspects on negative and positive influences.

The negative influences include the following aspects. Firstly, the presence of electricity consumption outlier data will reduce the accuracy of assessment. Then, the results of data mining cannot accurately reflect the characteristics of data. Finally, outlier data affect the judgement and decision of electric power system dispatcher [21]. It even threatens the safe operation of system. If outlier data cannot be correctly identified and effectively corrected, they will provide false prediction as a reference [35]. This affects the accuracy and reliability of prediction results. Furthermore, if outlier data are used as modeling data, it will interfere with the changing rules and influence the training accuracy. If outlier data are used as a predictor of test results, it can lead to erroneous judgement [28].

Outlier detection of electricity consumption data is to find out the relatively sparse and isolated outlier data patterns which are hidden in massive data [11]. In the early stage of data set preprocessing, we usually put outlier data as noise data. Although outlier detection finds the hidden data in the data set as the main purpose, outlier data mining is more valuable and meaningful than other types of mining [70]. Because one hundred thousand normal records are likely to cover only one rule, but ten outlier records are likely to cover ten different rules [28], [11]. Outlier detection of electricity consumption data may provide us more important information, so that we can find some real and unexpected knowledge, which can help us to understand the consumer behavior, capture the theft, find system vulnerabilities and failures and improve service quality [35] .

There is a growing interest in understanding how energy is spent in the commercial buildings. Furthermore, distributors of electricity want to know how to reduce the failure rate and detect anomalies which will serve the purpose of reliability and efficiency of the smart grid. In addition, they want to know how to visualize large volumes of energy data collected by power meters (sensors) in a building to find patterns, trends, and anomalies. In the end, our goal is to find how to automatically discover the anomaly, like unusual power consumption measurements highly differing from old observed patterns, and to reduce the energy cost of a building. For this task, anomalies are of special interest, because they can be caused either by faulty equipment or potentially misconfigured

devices consuming significantly more or less energy than required for proper operation [23].

Vulnerabilities in power systems can cause a major disaster considering the large usage of electricity in today's world. A widespread loss of power even for few days, could make devices ranging from cellphones to ATMs to traffic lights, and even lives to halt, if heating, air conditioning and health care systems exhaust their backup utilities. But when dealing with infrastructure that may even be system-critical, the number of failures must be reduced to an absolute minimum. Early signs of failures should be visible in power usage patterns [23]. By detecting anomalies in the usage patterns of electricity, machine learning algorithms can help utilities take customer care a step further, helping them identify homes at greatest risk of switching providers, listen to their concerns, and offer solutions. Our motivation therefore, will be concentrated on finding the appropriate information from the collected smart grids data and to employ suitable discovery method on these data to detect vulnerabilities and/or anomalies in the data.

Research questions and challenges

Information is necessary for energy players, and for decision-makers and regulators who must establish rules and mechanisms to provide a framework for a competitive market and allow smart grids to achieve optimal efficiency [12]. By 2020, the number of installed smart meters in Europe will reach 240 million while North America will have 150 million smart meters in use. China is forecasted to install about 400 million smart meters by that date. Japan would deploy about 60 million smart electricity meters and South Korea would plan to deploy between 500,000 and 1.5 million smart meters per year in homes before 2020. With the worldwide initiative of upgrading the traditional electrical grid to the Smart Grid, new challenges have been introduced, among which is the Big Data challenge. Many researchers are asking themselves what they can do to harness this deluge of data to make more intelligent operational and business decisions. With smart grids, it will be possible to obtain accurate, real-time data on consumption profiles and the status of electricity systems, with scope for isolating certain items on which it is appropriate to act (heating, household devices, etc.). Measures to promote energy efficiency and control consumption will benefit from the new behavioural data. The positive effect expected of smart grids is not necessarily a drop in prices but rather a reduction in the bills paid by consumers. It is estimated that gains from optimizing or reducing consumption will counterbalance higher prices due to regulation of and higher security on the electricity system[12]. Smart Grid generates a significant amount of data due to large scale deployment of digital technologies in distribution network on a daily basis, and these data must be processed in an efficient way in order to monitor the operational status of the grid. Research has shown that many existing techniques in the field of data analytics have not fully explored the value of data. However, serious attention has been given to this field and some well-known institutions have updated their teaching and research programs to

higher education students who can take up the challenges of big data research and applications in the near future. Data analytics competitors are competing to bring a set of IT tools and capabilities that are largely new to the utility industry [24].

While communications technology is seen as an essential enabling component of future smart grids, there are a number of challenges that must be addressed in order to have fully robust, secure and functional smart grid networks.

Regarding data analytics applications to the smart grid, the key challenges are focused on converting the tens of billions of data points coming from the millions of smart meters deployed around the country and turning them into actionable information for the grid operations. The data could generate information on how individual customers respond to requests of consumption reduction. It is also possible to use real-time metering data to discover unaccounted consumption when energy is being diverted and stolen. The aim is to improve grid reliability, outage response and reducing the cost of distribution operations. Occasionally emerged frauds or intrusions in smart grid systems have incurred significant loss when the suspicious activities were not detected or inefficiently processed. Therefore we found some of the research questions and challenges which needs to be answered to efficiently deploy the anomaly detection methods for electrical data:

1. **Availability of sufficient and valid input data:** For any machine learning algorithms to work and perform with good accuracy, is dependent on the size of input data. Sufficient amount of data will provide us with prior knowledge about the relationship between the variables, without which it will be difficult to generalize the unseen data. In our case of predicting unexpected events, working with small set of data could be challenging, as we may not capture required dependencies between the variables in the data, given the data is not large enough. We limit ourselves to only the size of the data in this discussion and won't be discussing about how the data is collected, assuming it is collected efficiently and systematically. It is important to know what is the data used for analysis, keeping in mind the purpose of the data and about the learning outcomes of it. Next, how many data points are enough to provide the evidence of the outcome achievement. Further details of how much data is enough can be found in [37]

2. **Dealing with unlabeled data:** The electrical data used for our analysis do not contain labels for anomalies and hence we use the general term "unlabeled data" as used in machine learning. We proceed by considering that most part of the data contain normal patterns and only a small part contain anomalous patterns. This assumption will be proved right or wrong later in the evaluation section after modeling and training the data. There is no explanation for each piece of data, it just contains data and nothing else. Challenge arises in determining which data to be considered as normal and which data as anomaly since we do not have labels

or classes to distinguish. Dealing with unlabeled data also challenges the choice of algorithm to be implemented which works the best with these data. Sometimes it can be taken into account that the accuracy in prediction will be high when using enough amount unlabeled data than labeled data [26].

3. **Identifying useful information:**The next challenge arises is feature selection. Identifying only the relevant features is important part of machine learning in anomaly detection. We need to take care of removing unnecessary data such as duplicates in the data which may be caused due to inefficient measurements and data containing 'null' values. Cleaning the data for removal of duplicates, null values and also to remove noise will help to analyse the data efficiently. Data Smoothing, by considering moving average windows readily available in python libraries can be used to do this cleaning process.

4. **Dimensionality Reduction:** Dimensionality reduction is used as an efficient technique when considering high dimensional datasets. Our anomaly detection process will be hard if we do not consider reducing the data to low dimension. To retain all the necessary information without losing any important aspects of the data will be challenging. For this we will employ well known dimensionality reduction machine learning technique such as PCA. PCA will reduce the dimensionality in the data considering only relevant information representing the whole data and ignoring the information which do not effect our anomaly detection process and which do not lose any important information about the data.

5. **Determining a suitable model of normality:** How to model the data will be the biggest challenge in our work. Modelling the data will determine the performance of the process of detecting the anomalies as well to describe the process efficiently. Model which best suit our data should be utilized. We will start with the basic and well known Gaussian Mixture Models and then PCA with residual parameters, followed by one of the efficient deep learning model, LSTM (Long Short-term Memory).

6. **Train the model:**Simply turning the work over to machines wont help: most machine learning and statistical mining techniques also hold the assumption that historical data, which is used to train the machine-learning model, behaves similarly to the target data, to which the model is later applied. We should be able to determine, what is the percentage of data to be used for training, validating and testing the model. The training data should be able to capture all the necessary features of the data which will be used to predict the future events. Choosing the training data will directly effect the accuracy of our model,

7. **Finding optimal model parameters:** Identify suitable statistical metric with which we will prove the correctness of our model. This statistical

metric should be able to calculate the threshold with which our model distinguishes between normal and abnormal patterns. This threshold value will the borderline between normal behavior and anomalous behavior. Choose a proper threshold above or below which data can be considered as anomaly/ies. Measures have to be taken care while fitting the model. A good model for the data should not overfit by using very large model parameters and it should not underfit as well using very less model parameters. Determining proper model parameters to include huge amount of data poses challenges to avoid the number of false positives and false negatives. False positives are data which are detected as anomalous are not truly anomalous and False negatives are data which are anomalous but our model did not detect.

8. **Deployment Efforts:** Great efforts have to be spent to develop more advanced and efficient algorithms for data analysis. Training the model with sufficient data, validating it and testing our methods consumed the most part of our work.

Contribution and outline

Background

A report on smart grid data management by analytic researchers at Accenture discusses five different data classes for smart grids in [5]: ¡Churn-rate¿ : in its broadest sense, is a measure of the number of individuals or items moving out of a collective group over a specific period of time. It is one of two primary factors that determine the steady-state level of customers a business will support.

1. **Operational data**: Represents the electrical behavior of the grid. It includes data such as voltage and current phasors, real and reactive power flows, demand response capacity, distributed energy capacity and power flows, and forecasts for any of these data items.

2. **Non-operational data**: Represents the condition, health and behavior of assets. It includes master data, data on power quality and reliability, asset stressors, utilization, and telemetry from instruments not directly associated with grid power delivery.

3. **Meter usage data**: Includes data on total power usage and demand values such as average, peak and time of day. It does not include data items such as voltages, power flows, power factor or power quality data, which are sourced at meters but fall into other data classes.

4. **Event message data**: Consists of asynchronous event messages from smart grid devices. It includes meter voltage loss/restoration messages, fault detection event messages and event outputs from various technical analytics. As this data is triggered by events, it tends to come in big bursts.

5. **Metadata**: Is the overarching data needed to organize and interpret all the other data classes. It includes data on grid connectivity, network addresses, point lists, calibration constants, normalizing factors, element naming and network parameters and protocols. Given this scope, managing metadata for a smart grid is a highly challenging task.

Outlier data have an important influence on accuracy, completeness and self-consistency of data quality. At the same time, it has important events information of electric power grid, such as power rationing, equipment failures and so on [45]. Thus, it is of great significance to identify, analyze and deal with outlier data. In residential or industrial electricity, it will produce large amounts of consumption data. Most consumption data are normal. However, there are also some outlier data. The generation of outlier data usually has the following reasons.

1. When electric power system is in operation, measurement and transmission processes of data acquisition system may generate outlier data. For example, outlier changes of data in the transmission are caused by the failure of data transmission system [43]. Thus, data may lose or confuse in the transmission process.

2. The data acquisition system is normal. But outlier changes of electric load are caused by special events, such as: line outage maintenance [14]. The occurrence of special events usually makes the electricity consumption data in a certain period of time as a null value. This is a kind of outlier phenomenon.

3. Significant electricity reforms make electricity consumption habits changed, such as: Ladder-type price [44]. Ladder-type price sets a base to residents, and more than the base price will increase the electricity price. This will lead to changes of electricity habits and result in outlier electricity consumption.

4. These conditions may also produce outlier data, such as: records errors, artificial forgery and so on. User's electricity consumption data may hide the user's electricity behavior habits. For data mining of electricity consumption data, on the one hand, it can understand the users demands for personalized service; on the other hand, it can detect the users stealing behavior and protect the benefits of enterprises.

Applying data mining techniques for power consumption data is a known approach for identifying abnormal usage behavior. Agarwal et al. examined 6 months of data from the UCSD campus, including aggregate power consumption of four buildings in [2]. Agarwal et al. focuses more on the setup of power meters and provide only simple visualization methods like line charts. Catterson et al. used an approach to monitor old power transformers in [8]. Their goal is to proactively search for abnormal behavior that may indicate the transformer is

about to fail. Similarly, McArthur et al. searched for anomalies to detect problems with power generation equipment in [30]. Jakkula and Cook compared several outlier detection methods to find which is better at identifying abnormal power consumption in [22]. J Seem used outlier detection to determine if the energy consumption for a day is significantly different from previous days energy consumption in [38]. This is a known approach for identifying abnormal system behavior.

There has been many work which already concentrates on securing the smart grid against the cyber- security threats. Rajasegarar et al. discusses anomaly detection techniques in wireless sensor networks considering the limited resources and their effects in [36]. Zhang et al. proposes a distribution intrusion detection system which developed and deployed intelligent and multiple analyzing modules in different layers of the smart grid in [46]. These systems uses support vector machines and Artificial immune systems to detect and classify possible cyber attacks. Fadlullah et al. and Gellings et al. have made steps towards machine to machine communications in [[15],[17]]. Fadlullah et al. discusses early warning detection technologies in smart grid communication using Gaussian process. With this warning system in place the smart grid control center can forecast the malicious activities to the smart grid which helps it to take measures against such activities. Perl et al. also proposes a solution to outlier detection using Gaussian Mixture model in [34]. Ozay et al. uses the machine learning algorithms to detect and classify the measurements as secure and non secure in [31]. It includes well known batch and online learning algorithms (supervised and semi-supervised). It takes advantage of the relationship between the statistical and geographical properties of the attack vectors. These properties are used to detect the unobservable patterns in the attacks. Janetzko et al. describes an analytical and visual approach in detecting anomalies and examining energy consumption data in [23]. The goal of the research was to enable the analyst to understand the power consumption behavior and to be aware of unexpected power consumption values.

For anomaly detection, machine learning techniques are often applied for detecting security threats, such as data injections or intrusions, instead of detecting bad data measurements. The latter is usually addressed by developing PMU placement algorithms that allow measurement redundancy [[10],[39]]. Hink et al. evaluated various machine learning methods to differentiate cyber-attacks from disturbances [20]. These methods include OneR, Naive Bayes, SVM, JRipper, and Adaboost. Among these approaches, the results show that applying both JRipper and Adaboost over a three-class space (attack, natural disturbance, and no event) is the most effective approach with the highest accuracy. Pan et al. propose a hybrid intrusion system based on common path mining [33]. This approach aggregates both synchrophasor measurement data and audit logs. In addition, SVMs have also been used in detection data cyberattacks. For instance, Landford et al. proposes a SVM based approach for detecting PMU data spoofs by learning correlations of data collected by PMUs which are electrically close to each other [25]. Due to the time sensitivity of event and anomaly detection on the Smart Grid, a major challenge of applying machine learning techniques

to perform these tasks is the efficiency of these approaches. This is especially important for online detection of events. Two different types of approaches have been used to address this challenge. Dimensionality reduction is a commonly used approach to improve learning efficiency by reducing the complexity of the data while preserving the most variance of the original data [41]. A significant amount of efforts have been made in this direction. For instance, Xie et al. propose a dimensionality reduction method based on principal component analysis for online PMU data processing [42]. It has been shown that this method can efficiently detect events at an early stage. A similar technique is used to reduce the dimensionality of synchrophasor data by extracting correlations of the data, and presenting it with their principal components in [13].

Understanding the data and Visualization

As machine learning techniques have become the de facto approaches for Big Data analytics, these techniques are also applied and adapted to process and analyze electrical data. A common objective of the electrical data analysis is to recognize and identify patterns, or signatures of events that occur on the Smart Grid. Therefore, a large number of these approaches are based on pattern recognition [4].Finding pattern or recognizing pattern can be considered as fitting a model in the data. Such an outcome is the result of a cascaded work process namely preparing data, searching pattern, evaluating knowledge and finally referring it towards the monitoring and control units. Data preparation includes data cleaning, data integration and transformation, data reduction and data warehousing which will explained in detail below. Before diving deep into understanding the data process-by-process, what we might learn from the dataset includes:

- The number of records in the dataset
- The number of attributes (or features)
- For each attribute, what are the datatypes(nominal, ordinal, or continuous)
- For nominal attributes, what are the values
- For continuous attributes, the statistics and distribution
- For any attributes, are there any missing values
- For any attributes, the number of distinct values

1. **Preparing Data:** Machine learning algorithms learn from data. It is critical that we feed them the right data for the problem we want to solve. Even if we have good data, we need to make sure that it is in a useful scale, format and even that meaningful features are included. This data preparation process has direct impact on the result, the more discipline we are in handling the data, the more consistent and better result we

will able to achieve. Below are the three important data processing steps widely followed when using machine learning algorithms, namely:

(a) **Step 1: Select Data**

(b) **Step 2: Preprocess Data**

(c) **Step 3: Transform Data**

2. **Searching Patterns:** Our next step in understanding the data is finding about the patterns in the data. Several statistical methods can be very helpful to find the patterns in the data which we will discuss below.

   (a) **Correlation:**

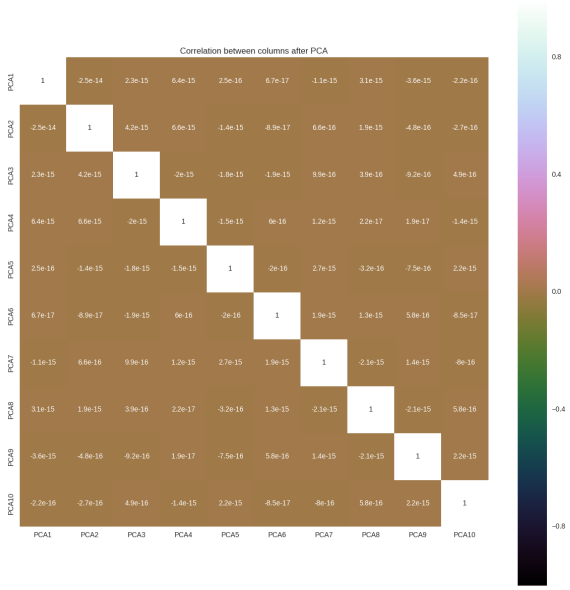

Figure 3: Correlation matrix of Data

3. **Knowledge Discovery:**

4. **Evaluating against the goal:**

In order to visualize the huge set of electrical data collected from electrical grids we had to store all the data in a database. Storing the data into the database

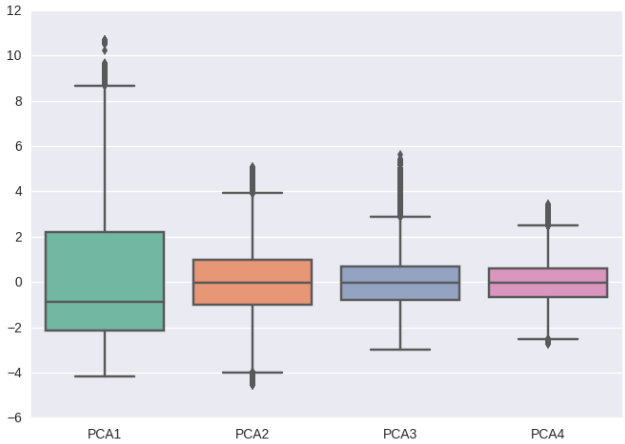

Figure 4: Box plot of Data

will not only allow us to compute the statistics better but also, will enable us to clearly visualize the patterns of the data in time windows such as weekday patterns, weekend patterns, hourly patterns and so on. Below we will discuss the structure of the data stored in the database:

Tablename: Energy-Data

Column Name	**Data Type**	**Default Value**	**Characteristics**	**Key**
Serial No	int	NOT NULL	AUTO-INCREMENT	Primary Key
Date	DATE	NOT NULL		
Time	TIME	NOT NULL		
V1	DOUBLE	NULL		
V2	DOUBLE	NULL		
V3	DOUBLE	NULL		
I1	DOUBLE	NULL		
I2	DOUBLE	NULL		
I3	DOUBLE	NULL		
I-N	DOUBLE	NULL		
Pges	DOUBLE	NULL		
Sges	DOUBLE	NULL		
CosPhi	DOUBLE	NULL		
Egy-trpt	DOUBLE	NULL		
Egy-con	DOUBLE	NULL		
Location	VARCHAR(50)	NOT NULL		

Data Description:

1. **Serial No**: This is the primary key to uniquely identify the record. This field will contain the serial numbers of the record. By record, while considering databases, we mean a row in the table Energy-Data.
2. **Date**: This field contains the date of the data. The value is stored in date format('YYYY-MM-DD').
3. **Time**: This field specifies the time when the data was collected. This value is stored as (hh:mm:ss).
4. **V1**: This is the voltage-1 of the 3-phase voltage expressed in Volts(V) which is stored as a double data type.
5. **V2**: This is the voltage-2 of the 3-phase voltage expressed in Volts(V) which is stored as a double data type.
6. **V3**: This is the voltage-3 of the 3-phase voltage expressed in Volts(V) which is stored as a double data type.
7. **I1**: This is the current in Phase-1 of the 3-Phase current expressed in Amperes(A)which is stored as a double data type.
8. **I2**: This is the current in Phase-2 of the 3-Phase current expressed in Amperes(A) which is stored as a double data type.
9. **I3**: This is the current in Phase-3 of the 3-Phase current expressed in Amperes(A) which is stored as a double data type.
10. **I-N**: This is the neutral current passing through the network which will be accounted for calculation of reactive power which is stored as a double data type.
11. **Pges**: Real power of the network, also known as working power expressed in units of Watts(W). It is stored as double datatype. Its the product of Sges and CosPhi values.
12. **Sges**: Actual power of the network, also known as apparent power calculated as a product of Voltage and Amperes expressed as Voltage-Ampere(VA). It is stored as double data type.
13. **CosPhi**: Power displacement, also known as phase displacement in the network. Sometimes referred to as Power factor which has no units, stored as double datatype.
14. **Egy-trpt**: Energy tranported in the network expressed in terms of Wattage-hour(Wh). Stored as double value.
15. **Egy-con**: Energy consumed at the receiver end as calculated by the smart meters. Expressed in terms of Wattage-hour(Wh), also stored as double datatype.

16. **Location**: New field added to the database which is not available in the csv files collected from the electrical grids. This field will be able to identify the data region which is very helpful for data analysis. Stored as string in the database.

Methodology: Anomaly Detection

Detecting and exploring of anomalies in time series is a very important aspect, especially when dealing with power consumption data of physical infrastructure. Abnormal behavior is defined as a difference from the expected normal pattern. We use the terms abnormal behavior, anomalies and outliers interchangebly referring to same definition of unusual data pattern. We introduce two methods of anomaly detection, both methods assume a daily power usage pattern which, of course, can be different for each day of the week. Both techniques are not limited to daily patterns, but can be easily adapted to the periodicity of the underlying data set. The first described method is based on a weighted prediction, where recent measurements have a higher impact than older measurements [23]. The latter approach is transforming the observed daily pattern in the frequency domain and looking for dissimilarity in a transformed space.

PCA-Based Anomaly Detection

PCA

Due to the high dimensionality nature of our dataset, we consider Principal Component Analysis(PCA) as one of the anomaly detection technique in our work. PCA is used to transform the high dimensionality datasets into low dimensional dataset. This low dimensional dataset captures all the necessary features by capturing the required variance which can be used to study the patterns exhibited by our data. Mathematically, PCA is an orthogonal linear transformation that transforms the data to new axes such that the first principal axes captures the greatest variance after projecting the data into this new axes, the second axes captures the second greatest variance, third axes the third greatest variance and so on. These axes are called the principal axes or the principal components(PCs). The principal components are ordered by the amount of variance that they capture. It is considered one of the reliable methods in capturing the correlation between the datasets. This descriptive method has been developed for the detection of linear relations between variables. However, if the relationships between the variables analyzed are not linear, the values of correlation coefficients can be lower. We consider normalizing our data before applying the PCA mainly because our data contains measures of different quantities such as voltages, current, power which are measured on different units. Normalizing the data is very essential preprocessing step before doing PCA, as the unnormalized data may lead to case where in each PC is dominated by a single variable. This would result in considering only one PC which has the largest variance and the remaining capturing very less variance. In effect the results of the analysis will

depend on what units of measurement are used to measure each variable. That would mean that principal component analysis should be done on raw data when all the measurements of our data are of same unit, which in turn will give more analysis weight on the variables which have higher variances. Therefore we standardize our data to mean equal to zero and standard deviation equal to one i.e $\mu = 0, \sigma = 1$. Standardizing the features so that they are centered around 0 with a standard deviation of 1 is not only important if we are comparing measurements that have different units, but it is also a general requirement for many machine learning algorithms.

PCA for Anomaly Detection:

Our set of electrical data collected from the smart grids contains many features. Considering all these features together can make our anomaly detection process complex. We therefore divide our dataset by constructing two subspaces based on the variances contributed by the features into normal subspace and residual subspace. Normal subspace contains most of the variance in the data as captured by the first 'K' principal components. Residual subspace is constructed by the remaining N-K components which contains very less variance in the data. The number of principal components(k) to choose will be discussed further in this section. Consider our data set containing i observations(rows) and j features(columns) which can be represented as a data matrix Y, so Y is a $i * j$ matrix. Each observation contains j features and denoted by single data point. From now on whenever we say a data point it represents i-th row containing j features of the data matrix Y.

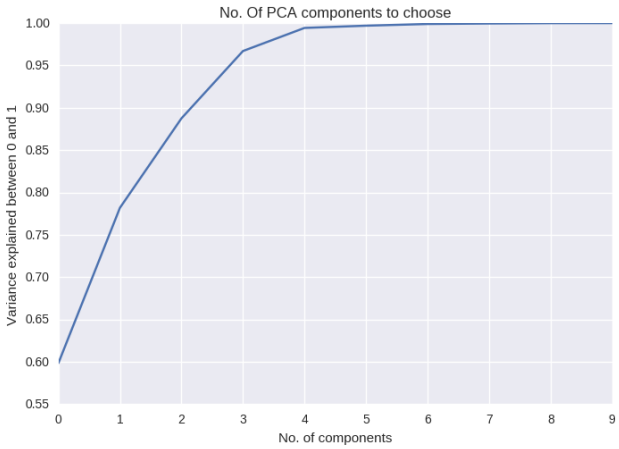

Figure 5: PCA Components

Prediction-Based Anomaly Detection

The basis for prediction is an observed pattern and the assumption that it is reoccurring (with slight modifications) in the future. We assume our data follows a regular underlying pattern and therefore also assume that the model describes the usual behavior well. Detecting anomalies using prediction follows this idea and is related to the statistical measure of residuals. Basically, this method predicts a value for each minute of the day by taking all previous measurement at the same time of the day. As an example, assume we predict the value for a Tuesday at 11:05am. We would now average all previous observed values of a Tuesday at 11:05am. Taking just an average would have the disadvantage of neglecting recent developments in the time series [23]. We therefore use a weighted averaging scheme with higher factors for recent values and linearly decreasing influence weights for older values. Further detailed explanations can be found in [19].

After predicting for each point in a time series the expected values based on all values occurring before this point in the time series, we can compute the difference between predicted and observed values. The difference is an indicator for the abnormality of the point in a time series but needs for higher expressiveness some kind of normalization. From the choice and the design of the prediction method we are assuming a model which may not be applicable to all observed time series. We counterbalance for this fact by calculating the average fitting of our model. More in detail, we compute the average deviation from the predicted values for the whole time series. If a whole time series is highly unpredictable, the differences between predicted and actual values are less meaningful compared to a case when a time series follows perfect daily patterns with small deviations. Computation of the anomaly score is summarized by the following equation [23]:

$$\text{anomaly[time]} = \frac{|predVal[time] - obsVal[time]|}{avgtTime(|predVal[t] - obsVal[t]|)}$$

Clustering-Based Anomaly Detection

The second approach for detecting anomalies in time series data is similarity-based. We assume often-observed patterns to be the usual behavior and rarely occurring patterns to be abnormal. Following this idea, we first have to define and compute the similarity of patterns in order to detect whether a pattern occurs more than once. The approach described in this section is proposed and presented by Bellala et al. in [[6],[7]]. The time series is first partitioned into days and afterwards transformed by a Fourier transformation into the frequency domain. Each day of the time series is resulting in a k-dimensional vector in the frequency domain with k being a parameter of the transformation process. The next step described by Bellala et al. is a dimension reduction by multi-dimensional scaling into a two-dimensional space. The density distribution in the reduced MDS space is now interpreted as an anomaly score. Points (time series of a single day) being in a high-density area with many (similar) neighbors are assumed to reflect the usual behavior. Outliers in the 2D space can be seen as

days with unusual values and are assigned a high anomaly score. This technique only takes the frequency domain into account and does not integrate external effects like weather data or week of the day.

Gaussian Mixture Model-Based Anomaly Detection

Here we propose yet another unsupervised algorithm for anomaly detection using Gaussian Mixture Model. The Gaussian Mixture Model(GMM) is used to model the probability density function of a feature vector, x, by the weighted combination of M multi-variate Gaussian densities (λ) [29] :

$$p(x|\lambda) = \sum_{i=1}^{M} p_i g_i(x) \tag{1}$$

First we fit a Gaussian Mixture Model (GMM) with Gaussians centered at each data point to a given data set. We then estimate the mixture proportion by applying Expectation Maximization algorithm to the given dataset. Each mixture proportion will represent the degree to which the point is a cluster center. The higher the mixture proportion, the more likely it is a cluster center, which means it has higher influence on other points. Reversely, the lower the mixture proportion is, the less likely the point is a cluster center, which means it has lower influence on other points. The outlier factor at each data point is then defined as a weighted sum of the mixture proportions with weights representing the similarities to other data points.

Given a set of data points$X = x_1, ..., x_n$, a standard Gaussian mixture model clustering seeks to maximize the scaled log-likelihood function [34]

$$l(\pi_{1:m}, \mu_{1:m}, \lambda; X) = \frac{1}{n} \sum_{i=1}^{n} log[\sum_{j=1}^{m} \pi_j p(x_i|\mu_j, \lambda)] \tag{2}$$

where m is the number of model components, $\pi_j = p(\omega_j|\lambda)$ represents the strength of jth component ω_j with $\sum_{i=1}^{m} \pi_i = 1$ and $\pi_{1:m}$ is a vector composed of π_j for j = 1, . . . , m. The probability $p(x_i|\mu_j, \lambda)$ is a Gaussian and λ is a vector of parameters specified below.

In the standard mixture model μ_j is the unknown mean vector for jth component and is estimated together with other parameters using an EM algorithm. Since our goal is not clustering but an estimation of an outlier factor at every data point, we assume that each data point is a cluster center. Thus, we set, m = n and $\mu_j = x_j$ for j = 1, . . . , n. This way, the mixture proportion π_j represents the likelihood of point x_j to be a cluster center. We obtain a simplified version of Eq. (2):

$$l(\pi_{1:m}; X) = \frac{1}{n} \sum_{i=1}^{n} log[\sum_{j=1}^{n} \pi_j p(x_i|x_j, \lambda)] \tag{3}$$

In E-step we compute for each class i = 1, . . . , n and for each data point k = 1, . . . , n:

$$p(x_i|x_k, \lambda_t) = \frac{p(x_k|x_i, \lambda_t)p(x_i|\lambda_t)}{p(x_k|\lambda_t)} = \frac{p(x_k|x_i)\pi_i(t)}{\sum_{j=1}^{n} p(x_k|x_j)\pi_j(t)} \tag{4}$$

Our M-step is particularly simple, since we only need to update the mixture proportions:

$$\pi_j(t+1) = \frac{1}{n}\sum_{k=1}^{n} p(x_i|x_k, \lambda_t) \tag{5}$$

Plugging in Eq. (4) in Eq. (5) gives

$$\pi_j(t+1) = \frac{1}{n}\sum_{k=1}^{n} \frac{p(x_k|x_i)\pi_i(t)}{\sum_{j=1}^{n} p(x_k|x_j)\pi_j(t)} \tag{6}$$

the term $p(x_k|x_j)\pi_j(t)$ represents how much point x_k is influenced by point x_j with $p(x_k|x_j)$ being the strength of the connection and $\pi_j(t)$ measuring the importance of point j. This motivates the proposed definition of the outlier factor at point x_k as

$$F_k = \sum_{j=1}^{n} p(x_k|x_j)\pi_j(t) \tag{7}$$

According to Eq. (7), the smaller F_k, the more likely is data point x_k to be an outlier.

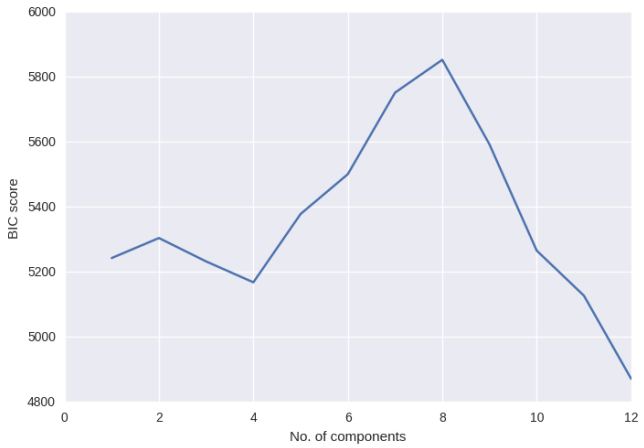

Figure 6: BIC score to determine the number of gaussians to choose

Evaluation:

In order to evaluate the performance and correctness of GMM, PCA-based and LSTM based anomaly detection we have trained and evaluated the implemented models based on the set of data provided by "PolyEnergyNet Resiliente Polynetze zur sicheren Energieversorgung". We will perform the analysis by grouping the datasets into two groups namely, weekday and weekends in order to identify the patterns of the weekday and weekends separately. We believe the patterns differ significantly between the weekday and weekends and hence have divided the analysis into two parts.**Find out if there are any specific methods to split the dataset into training and testing, particularly for time series data or our assumption of taking one week data to train and another week data to test is correct ?** Firstly, we use one week's weekday data initially and split randomly into 80 percent training data and remaining 20 percent as the test data. This percentage split is just random considering that the training dataset should be more than the testing dataset. We can choose anything between 50-50 split to 60-40, 70-30, or 80-20 depending on the size of the dataset. Since we have taken one week's data we want to capture most of the statistical values and hence decided to go with 80-20 split. We would also look into evaluating against 70-30 split if time permits or if the results are not as expected. For this we use the sklearn libraries model_selection.train_test_split() function which takes data and percentage variable (which should be between 0 and 1) as parameters. The percentage variable if set to 0.5 gives us 50 percent training and 50 percent testing split of the data. Similarly, if the variable takes 0.6 then its a 60-40 split and so on. This data has been preprocessed to take care of missing data values and considering only the relevant columns as features for our anomaly detection purpose.

Conclusion

Bibliography

[1] ABB. What is a smart grid?, 2016. [Online; accessed 2016-07-22].

[2] Yuvraj Agarwal, Thomas Weng, and Rajesh K Gupta. The energy dashboard: improving the visibility of energy consumption at a campus-wide scale. In *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 55–60. ACM, 2009.

[3] Marco Aiello and Giuliano Andrea Pagani. The smart grid's data generating potentials. In *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*, pages 9–16. IEEE, 2014.

[4] Miftah Al Karim, Moustafa Chenine, Kun Zhu, Lars Nordstrom, and Lars Nordström. Synchrophasor-based data mining for power system fault analysis. In *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pages 1–8. IEEE, 2012.

[5] A Analytics. Achieving high performance in smart grid data management: Making sense of the data deluge. Technical report, Accenture Analytics, Tech. Rep, 2010.

[6] Gowtham Bellala, Manish Marwah, Martin Arlitt, Geoff Lyon, and Cullen Bash. Following the electrons: methods for power management in commercial buildings. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 994–1002. ACM, 2012.

[7] Gowtham Bellala, Manish Marwah, Martin Arlitt, Geoff Lyon, and Cullen E Bash. Towards an understanding of campus-scale power consumption. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 73–78. ACM, 2011.

[8] Victoria M Catterson, Stephen DJ McArthur, and Graham Moss. Online conditional anomaly detection in multivariate data for transformer monitoring. *IEEE Transactions on Power Delivery*, 25(4):2556–2564, 2010.

[9] Shing-Chow Chan, Kai Man Tsui, HC Wu, Yunhe Hou, Yik-Chung Wu, and Felix F Wu. Load/price forecasting and managing demand response

for smart grids: Methodologies and challenges. *IEEE Signal Processing Magazine*, 29(5):68–85, 2012.

[10] Jian Chen and Ali Abur. Placement of pmus to enable bad data detection in state estimation. *IEEE Transactions on Power Systems*, 21(4):1608–1615, 2006.

[11] Wen Chen, Kaile Zhou, Shanlin Yang, and Cheng Wu. Data quality of electricity consumption data in a smart grid environment. *Renewable and Sustainable Energy Reviews*, 2016.

[12] Cédric Clastres. Smart grids: Another step towards competition, energy security and climate change objectives. *Energy Policy*, 39(9):5399–5408, 2011.

[13] Nischal Dahal, Roger L King, and Vahid Madani. Online dimension reduction of synchrophasor data. In *Transmission and Distribution Conference and Exposition (T&D), 2012 IEEE PES*, pages 1–7. IEEE, 2012.

[14] Tamraparni Dasu and Theodore Johnson. Hunting of the snark: Finding data glitches using data mining methods. In *IQ*, pages 89–98. Citeseer, 1999.

[15] Zubair Md Fadlullah, Mostafa M Fouda, Nei Kato, Xuemin Shen, and Yousuke Nozaki. An early warning system against malicious activities for smart grid communications. *IEEE Network*, 25(5):50–55, 2011.

[16] David M Gann, Mark Dodgson, and Dheeraj Bhardwaj. Physical–digital integration in city infrastructure. *IBM Journal of Research and Development*, 55(1.2):8–1, 2011.

[17] Clark W Gellings. *The smart grid: enabling energy efficiency and demand response*. The Fairmont Press, Inc., 2009.

[18] Vehbi C Gungor, Dilan Sahin, Taskin Kocak, Salih Ergut, Concettina Buccella, Carlo Cecati, and Gerhard P Hancke. Smart grid technologies: Communication technologies and standards. *IEEE transactions on Industrial informatics*, 7(4):529–539, 2011.

[19] Ming C Hao, Halldor Janetzko, Sebastian Mittelstädt, Water Hill, Umeshwar Dayal, Daniel A Keim, Manish Marwah, and Ratnesh K Sharma. A visual analytics approach for peak-preserving prediction of large seasonal time series. In *Computer Graphics Forum*, volume 30, pages 691–700. Wiley Online Library, 2011.

[20] Raymond C Borges Hink, Justin M Beaver, Mark A Buckner, Tommy Morris, Uttam Adhikari, and Shengyi Pan. Machine learning for power system disturbance and cyber-attack discrimination. In *Resilient Control Systems (ISRCS), 2014 7th International Symposium on*, pages 1–8. IEEE, 2014.

[21] Shyh-Jier Huang, Jeu-Min Lin, et al. Enhancement of anomalous data mining in power system predicting-aided state estimation. *IEEE Transactions on Power Systems Pwrs*, 19(1):610–619, 2004.

[22] Vikramaditya Jakkula and Diane Cook. Outlier detection in smart environment structured power datasets. In *Intelligent Environments (IE), 2010 Sixth International Conference on*, pages 29–33. IEEE, 2010.

[23] Halldór Janetzko, Florian Stoffel, Sebastian Mittelstädt, and Daniel A Keim. Anomaly detection for visual analytics of power consumption data. *Computers & Graphics*, 38:27–37, 2014.

[24] CS Lai and Malcolm D McCulloch. Big data analytics for smart grid. *Newsletter*, 2015.

[25] Jordan Landford, Rich Meier, Richard Barella, Xinghui Zhao, Eduardo Cotilla-Sanchez, Robert B Bass, and Scott Wallace. Fast sequence component analysis for attack detection in synchrophasor networks. *arXiv preprint arXiv:1509.05086*, 2015.

[26] Feng Liang, Sayan Mukherjee, and Mike West. The use of unlabeled data in predictive modeling. *Statistical Science*, pages 189–205, 2007.

[27] Jing Liu, Yang Xiao, Shuhui Li, Wei Liang, and CL Philip Chen. Cyber security and privacy issues in smart grids. *IEEE Communications Surveys & Tutorials*, 14(4):981–997, 2012.

[28] Guojun Mao, Li-Juan Duan, Shi Wang, and Yun Shi. Principle and algorithm of data mining, 2005.

[29] Simone Marinai and Hiromichi Fujisawa. *Machine learning in document analysis and recognition*, volume 90. Springer, 2007.

[30] Stephen DJ McArthur, Campbell D Booth, JR McDonald, and Ian T McFadyen. An agent-based anomaly detection architecture for condition monitoring. *IEEE Transactions on Power Systems*, 20(4):1675–1682, 2005.

[31] Mete Ozay, Inaki Esnaola, Fatos Tunay Yarman Vural, Sanjeev R Kulkarni, and H Vincent Poor. Machine learning methods for attack detection in the smart grid. 2015.

[32] Ben Packer. 7 reasons why utilities should be using machine learning, 2015. [Online; accessed 2016-09-15].

[33] Shengyi Pan, Thomas Morris, and Uttam Adhikari. Developing a hybrid intrusion detection system using data mining for power systems. *IEEE Transactions on Smart Grid*, 6(6):3104–3113, 2015.

[34] Y Perl and Amir Pnueli. Outlier detection with globally optimal exemplar-based gmm. 2009.

[35] Duc-Son Pham, Svetha Venkatesh, Mihai Lazarescu, and Saha Budhaditya. Anomaly detection in large-scale data stream networks. *Data Mining and Knowledge Discovery*, 28(1):145–189, 2014.

[36] Sutharshan Rajasegarar, Christopher Leckie, and Marimuthu Palaniswami. Anomaly detection in wireless sensor networks. *IEEE Wireless Communications*, 15(4):34–40, 2008.

[37] Gloria Rogers. Death by assessment: How much data are too much? *Communications Link*, 2002.

[38] John E Seem. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and Buildings*, 39(1):52–58, 2007.

[39] Aditya Tarali and Ali Abur. Bad data detection in two-stage state estimation using phasor measurements. In *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pages 1–8. IEEE, 2012.

[40] Michael Troiano. Internet of things drives growth for smart grid innovation, 2007-2016. [Online; accessed 2016-07-22].

[41] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.

[42] Le Xie, Yang Chen, and PR Kumar. Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis. *IEEE Transactions on Power Systems*, 29(6):2784–2794, 2014.

[43] Man Xu, Zongxiang Lu, Ying Qiao, Ningbo Wang, and Shiyuan Zhou. Application of change-point analysis to abnormal wind power data detection. In *2014 IEEE PES General Meeting— Conference & Exposition*, pages 1–5. IEEE, 2014.

[44] Yang Yu, Yingjie Li, and Jeff Heflin. Detecting abnormal semantic web data using semantic dependency. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 154–157. IEEE, 2011.

[45] Jun Zhang and Hong Wang. A new pretreatment approach of eliminating abnormal data in discrete time series. In *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05.*, volume 1, pages 4–pp. IEEE, 2005.

[46] Yichi Zhang, Lingfeng Wang, Weiqing Sun, Robert C Green II, and Mansoor Alam. Distributed intrusion detection system in a multi-layer network architecture of smart grids. *IEEE Transactions on Smart Grid*, 2(4):796–808, 2011.