2022

Data-Enabled Distribution Grid Management

Zohreh Sadat Hosseini

DATA-ENABLED DISTRIBUTION GRID MANAGEMENT

A Dissertation

Presented to

the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Zohreh Sadat Hosseini

March 2022

Advisor: Dr. Amin Khodaei

Author: Zohreh Sadat Hosseini
Title: DATA-ENABLED DISTRIBUTION GRID MANAGEMENT
Advisor: Dr. Amin Khodaei
Degree Date: March 2022

Abstract

In 2020, U.S. electric utilities installed more than 94 million advanced meters, which brought the percentage of residential customers equipped with smart meters to 75%. This significant investment allows collecting extensive customer data at the distribution level, however, the data are not currently leveraged effectively to help with system operations. This dissertation aims to use the smart meters' data to improve the grid's reliability, stability, and controllability by solving two of the most challenging problems at the distribution level, namely distribution network phase identification and outage identification.

Distribution networks have typically been the least observable and most dynamic and locally controlled elements in the power grid. Complete information about the network topology is continuously changing and is not always readily available when needed. Lack of phase connectivity information is a challenge, especially when rebalancing the grid and also in the aftermath of outages caused by extreme events. Traditionally, phase identification is executed manually. In this dissertation, a machine learning-based data mining method for accurate and efficient phase identification of residential customers is proposed by leveraging power consumption data collected through smart meters. The proposed method uses a high-pass filter to remove the redundant and irrelevant segments of the power consumption time series, and accordingly identifies the residential customers' phase connectivity through a modified clustering algorithm.

Accurate connectivity information among customers is essential for outage identification and management in distribution networks. Extreme weather events can cause significant damage to electric power grid infrastructure and lead to widespread power outages. The frequency and the intensity of these events are continuously increasing as a direct result of climate change. Identifying grid components that are damaged is the first step to recovering from extreme weather-related power outages. An effective data mining method in identifying distribution network line outages is presented in this dissertation by leveraging data collected through AMI. The line outage identification method is developed based on a Multi-Label Support Vector Machine (ML-SVM) classification scheme that utilizes the status of customers' smart meters as input data and identifies the outage/operational status of distribution lines.

Numerical simulations demonstrate the effectiveness of the proposed models and their respective viability in achieving the targeted operational objectives.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Amin Khodaei, for his academic supervision, motivation, and immense knowledge during my PhD study and research. It has been a genuine pleasure to work with him, and I am truly grateful for his consistent supports.

I want to greatly thank Aleksi Paaso, director of distribution planning, smart grid, and innovation at ComEd for his kind guidance and advice during my PhD research.

My gratitude extends to my PhD defense committee members, Dr. Rui Fan, Dr. Mohammad Matin, and Dr. Ryan Elmore for taking the time to review my dissertation and providing invaluable feedbacks.

My deep appreciation also goes to my inspiring and supportive mother, my dearest deceased father, and my siblings for their abundant care and encouragements throughout my life.

A very special thanks to my husband, Mohsen whose tremendous love, understanding, and encouragement keep me motivated and confident. Working with him as a lab mate surely made my research both productive and enjoyable.

Finally, I would like to thank all my friends and colleagues at the University of Denver for the cherished time spent together at KLab.

Table of Contents

List of Figures

List of Tables

Nomenclature

Chapter Two:

Indices

t	Index for time
i	Index for clusters/phases
k	Index for clusters/phases
j	Index for residential customers
n	Index for residential customers

Parameters

K	Number of clusters/phases
N	Total number of residential customers

Variables

x	object/residential customer
S	Aggregated power consumption of each phase at the substation
P	Power consumption time series of each residential customer
M	Power consumption time series *measurement*
C	Centroid of each cluster
L	Assigned cluster to each customer
D	Distance between residential customers and centroids

Chapter Three:

Indices

i	Index for training data
j	Index for training data
d	Index for smart meters
l	Index for lines

Sets

D	SVM Training set

Parameters

w	SVM model parameter
b	SVM model parameter
B	Number of smart meters
N	Number of training data
L	Number of classes/lines
c	Tradeoff parameter
y_p	Positive parameter for polynomial kernel
y_r	Positive parameter for RBF kernel
r	Nonnegative parameter for polynomial kernel
d’	Degree of polynomial kernel
β	$F_{\beta\text{-score}}$ parameter

Variables

x	Outage status of each smart meter
y	Status of each line
ξ	Slack variable

1 Chapter One: Introduction

Utilities are upgrading their networks from the manually-read analog meters to new technologies such as advanced metering infrastructure (AMI) or smart meters, which provide a great opportunity for increasing observability and controllability of the distribution network. More than 94 million smart meters were installed in the U.S. in 2020, in which 88% are smart residential meters [1]. Fig. 1.1 shows the percentage of installed smart meters in the US in 2021. Although extensive data is collected at the distribution level, they are not used effectively in advanced applications to support system operations.

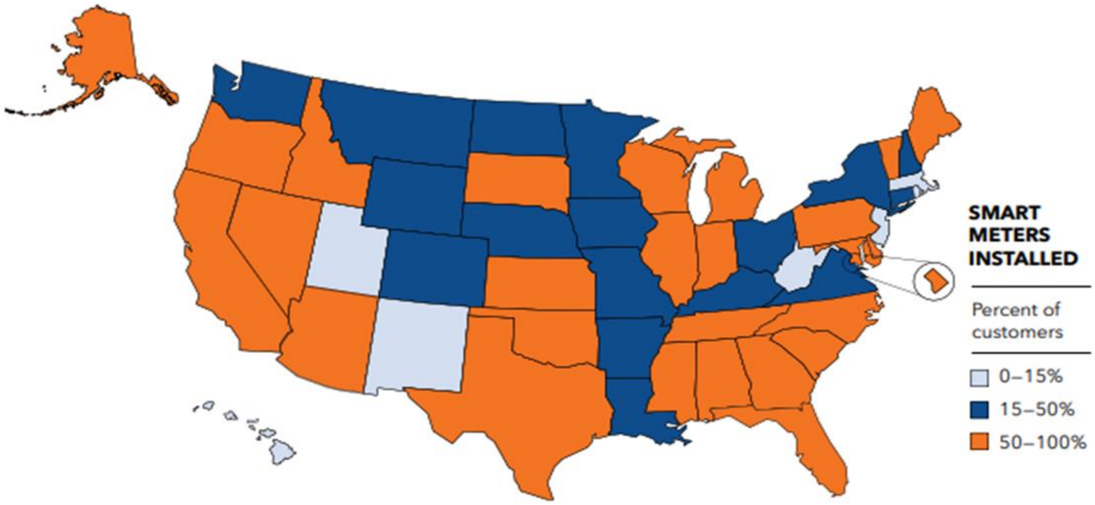


Fig. 1.1: Smart meter deployment in the US in 2021 [2].

Smart meters measure and record customers' electricity usage frequently and at given intervals (every 15 min, 30 min, or 1 h). This dissertation aims to use the smart meters’ data and provide a solution for two of the most challenging problems at the distribution level, which are distribution network phase identification and outage identification. Utilities can leverage the proposed solutions to improve the reliability, stability, and controllability of the distribution grid.

Distribution networks have typically been the least observable and most dynamic and locally controlled element in the power grid. Complete information about the network topology is continuously changing and is not always readily available when needed.

A distribution network is used to deliver electricity at medium voltages to end-use customers spread over vast geographical areas. Electricity is commonly generated as 3-phase AC and is injected into the high voltage transmission network, which is subsequently stepped down to be delivered to the distribution network. Residential customers are mostly single-phase and are distributed across the network. Automated monitoring and control in the grid have traditionally been mostly deployed in the transmission networks. At the distribution level, however, there are not as many integrated solutions, and the monitoring has been limited to distribution automation (DA) and, most recently, advanced metering infrastructure (AMI). As the distribution network is the least observable and most dynamic and locally controlled element in the grid, utilities may not have access to reliable and complete updated information about the distribution network, such as the phase to which each customer is connected [1]. Lack of phase connectivity information is a challenge, especially when it comes to rebalancing the grid and outage identification and management.

Phase identification and network rebalancing are a hard, costly, and time-consuming task for electric utilities; however, it is of great importance to future grid planning and advanced distribution management system (ADMS) as well as outage management system (OMS) type operations.

As a part of distribution planning, electric utilities annually review phase balancing of all the feeders under peak load conditions. However, even a balanced network would become unbalanced over time [3] because of the addition of new customers, maintenance, restoration, reconfiguration, and change in customers' consumption patterns. For instance, after weather events such as storms, customers will be connected to a phase that leads to the fastest restoration, thus potentially distorting the phase balance. Unbalanced loading can cause several problems for the network, including but not limited to poor power quality such as unbalanced service voltages and over/under voltages, increased power losses, reducing the lifetime of grid assets like transformers, overheating, reduced distributed generation hosting capacity [4]-[6], and delayed power restoration and subsequently longer outage-time [7]. Additionally, accurate connectivity information is required for efficient renewable energy sources' integration [8], [9], loss reduction, operational improvements, and rebalancing [5], as well as outage identification and management in low voltage distribution networks [9].

Phase identification is traditionally executed manually, although there are existing voltage measurement-based methods that are not always reliable.

Phase identification is traditionally implemented either manually or through signal injection [10]-[11], both of which are costly, labor-intensive, time-consuming, and errorprone. In addition, wrong phase identification causes errors in topology detection, state estimation, and fault location detection [1].

Accurate connectivity information among consumers as a part of topology identification is essential for outage identification and management in low voltage distribution networks.

Extreme weather events can cause significant damage to electric power grid infrastructure and lead to widespread power outages. The frequency and the intensity of these events is continuously increasing as a direct result of climate change [12]. From 2002 to 2015, extreme weather events have caused more than 87% of major power outages involving 50,000 or more customers in the U.S. [13].

The identification of grid components that are damaged is the first step to recovering from extreme weather-related power outages. These components can be easily identified in the generation and transmission levels; however, this is not the case for distribution level components. Traditionally, the distribution grid has not been fully observable to grid operators, causing outage location identification a challenging task. This task is currently performed by investigating the feeder configuration map and the protection design manual to identify the overall outage locations [14], i.e., an expert-experience-based method. Although the expert-experience-based method may be able to achieve highly accurate solutions, it is proved to be laborious, costly, and time-consuming [15], [16]. Time is of particular essence in this case. As the outage duration increases, the associated outage cost will increase almost exponentially. The estimated outage cost for residential customers during interruptions of 1 min, 20 min, 1 h, 4 h and 8 h is quantified as 0.001, 0.093, 0.482, 4.914, and 15.690 $/kW, respectively [17]. An automated method with a short computation time would be significantly useful in this case, especially for larger networks.

In this dissertation, a machine learning-based data mining method for accurately and efficiently identifying the phase of each residential customer in a distribution network is proposed. The proposed method leverages power consumption data collected through the AMI and uses a high-pass filter to remove the redundant and irrelevant parts of the power consumption time series. It then identifies the residential customers' phase connectivity by proposing a modified clustering algorithm.

In addition, to solve the outage identification problem, a machine learning-based data mining method is presented to quickly and efficiently identify distribution lines outages in response to extreme events and by leveraging smart meter data collected through AMI.

1.1 Dissertation Overview

The main body of this dissertation is based on the collection of articles published during the Ph.D. studies. The rest of this dissertation is organized as follows.

Chapter 2 focuses on distribution phase identification as one of the most challenging problems in distribution networks. At first, the existing literature in phase identification is reviewed and the problem statement and phase identification model outline are presented. The effectiveness of the proposed method is investigated under complete and incomplete data scenarios as well as in the presence of residential solar PVs and a brief discussion based on the cases studied concludes this chapter.

Chapter 3 focuses on distribution outage identification and presents a multi-label support vector machine (ML-SVM) scheme to identify distribution lines outages in response to extreme weather events by leveraging AMI data. The effective and acceptable performance of the proposed scheme is validated through numerical simulations for both small and relatively large test systems. The accuracy of the proposed method is then examined in case of lost last gasp signals in the distribution system. Utility companies can reap the benefits of this intelligent method to accelerate the process of grid response and recovery and consequently, decrease the associated outage durations and costs.

2 Chapter Two: Distribution Phase Identification

2.1 Introduction

Distribution networks, as the last element of a power system, are used to deliver electricity at medium voltages to end-use customers spread over vast geographical areas. Fig. 2.1 demonstrates a big picture of a power system. Electricity is commonly generated as 3-phase AC and is injected to the high voltage transmission network, which is subsequently stepped down to be delivered to the distribution network, where residential customers are mostly single-phase and are distributed across the network.

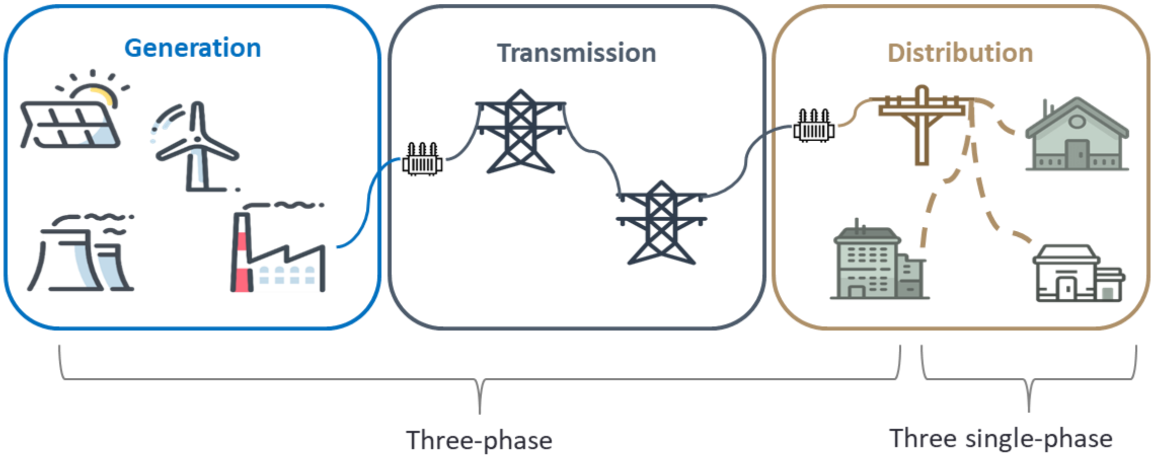


Fig. 2.1: A big picture of a power system containing generation, transmission, and distribution networks.

At the distribution level, there are not as many integrated solutions for automated monitoring and control and the monitoring has been limited to distribution automation

(DA) and, most recently, AMI. As the distribution network is the least observable and most dynamic and locally controlled element in the grid, utilities may not have access to reliable and complete updated information about the distribution network, such as the phase to which each customer is connected [1]. However, accurate connectivity information among consumers as a part of topology identification is essential for efficient renewable energy sources' integration [7], [8], loss reduction, operational improvements, and rebalancing [5], as well as outage identification and management in low voltage distribution networks [9]. This dissertation aims to propose an efficient data-driven method to identify phase connectivity of residential customers in distribution networks.

2.2 Literature Review

The literature available for phase identification is limited. In general, two common methods are available to identify the phase connectivity of residential customers, including hardware-based methods and software-based methods. Hardware-based methods focus on using specially-designed devices [18], PLC smart meters [19], signal injection methods [10]-[11], and micro-synchrophasors [1]. Hardware-based methods are costly because of the additional devices that need to be installed in the network, as well as the cost of labor to deploy such devices. On the other hand, software-based methods are becoming more popular due to the increasing deployment of AMI. One of the most popular methods to identify the phase that each customer is connected to is to use the time series of voltage measurements. It is shown that the time series of voltage measurement of each residential customer has the highest correlation to the voltage series of the connected phase at its respective substation. Based on this, a study in [1] proposes an algorithm to analyze crosscorrelation coefficients over-voltage magnitudes by considering phase angle differences on different phases. The proposed algorithm uses data from high-precision phasor measurement units (micro-synchrophasors or uPMUs) to compare correlations of the residential customers with measured voltage at the substation for each phase. This method may, however, be impractical for smart meters due to the significant difference in interval reads of smart meters (30 min) versus uPMUs’ measurements (potentially hundreds per cycle). Similarly, authors in [3] and [20]-[23] use correlation of voltage measurements to identify phase connectivity. The study in [24] uses linear regression and correlation on voltage measurements as well as kilowatt-hour measurements from AMI to estimate secondary connectivity and primary-side voltage profiles. There are a few studies that use clustering methods and voltage measurements for phase identification. For instance, based on the voltage correlation, a constraint-driven hybrid clustering (CHC) algorithm [25], k-means clustering [23], spectral clustering [26], and a combined feature-based clustering approach with principal component analysis (PCA) [27] are developed to identify the phase that each residential customer is connected to. Other approaches which use voltage measurements are also available in the literature, such as using the Tabu search method [28] for estimating the phasing of laterals by using circuit measurements and load flow information.

The lack of adequate historical voltage measurements can emerge as a major disadvantage of using voltage-based methods. In this respect, a few studies focus on using power measurements instead of voltage measurements for phase identification. For instance, the study in [29] uses an integer programming and branch and bound search algorithm on the power measurements collected from residential customers to identify residential customer-connected phases. As this method uses integer programming to model the problem, the number of customers should be known. Authors in [30] use similar setup and assumptions as to the previous study. The power measurements are used to set up a system of linear equations based on the principle of conservation of energy. These equations are then analyzed to estimate a tree network that optimally fits the power measurements. This method needs the number of customers to optimize the number of required measurements. The study in [31] uses grouped smart meter power data to detect mixed-phase groups in a case that individual smart meter data for customers is not available. Authors in [32] also use power measurements and present a spectral and saliency analysis for phase identification. The proposed method needs more than one month of customers' data. In addition, this method requires other customers' data. In other words, by decreasing the smart meter penetration in the network, i.e., losing the data of a portion of customers, the accuracy of the proposed method would decrease. The study in [33] proposes a data-driven approach based on PCA and its graph-theoretic interpretations. In addition to requiring a number of connected customers to each phase, the proposed method is not able to identify phase connectivity in the presence of unmetered loads in the network.

The proposed method in this dissertation identifies the phase connectivity of the residential customers by leveraging data collected through smart meters. The phase identification method is developed based on a combination of filtering and a modified clustering algorithm that utilizes the power consumption time series of customers. It is further made sure that the proposed method works effectively for scenarios with incomplete data [34]. The advantages and contributions of the proposed method compared to the existing methods are summarized as follows:

- No additional hardware, monitoring tools, or signal injection devices are required;
- Unlike the existing methods, the proposed method does not need to collect voltage and current measurements; instead, the power consumption profiles, which are recorded and kept by the utility, are used;
- The number of total customers can be unknown, which enables partial phase identification for any selected part of the network. In other words, the number of total customers is not a decisive factor in this method;
- Using the proposed method, the phase connectivity of each customer is individually identifiable without requiring other customers' data;
- The proposed method is unsupervised, so a labeled dataset is not required;
- The topology of the network is not required to be known;
- As the proposed method applies a preprocessing step to the dataset for removing redundant parts of the power measurement time series, the length of the required time series is significantly shortened;
- The proposed method is highly efficient in terms of the computation time;
- As the proposed method is applicable for each individual residential customer, it is robust against the unmetered loads in the network;
- The proposed method is applicable to both balanced and unbalanced networks;

- The proposed method can identify residential customers' phase connectivity in networks consisting of both three-phase and single-phase nodes; and
- The proposed method can identify residential customers' phase connectivity under incomplete data scenarios.

2.3 Model Outline and Formulation

Consider a set of N single-phase residential customers in a distribution feeder. The phase that each residential customer is connected to is unknown. The number of residential customers that are connected to each phase is also unknown. Time series of residential customers' power consumptions are recorded frequently and at given intervals (commonly every 15 min, 30 min, or 1 h) [35]. In the same fashion, the aggregated power consumptions are recorded for each phase at the associated substation. A two-step preprocessing/clustering approach is proposed to identify the phase connectivity of the residential customers in the feeder. In the first step, by applying a high-pass filter to the residential customers' power consumption time series, as well as the phases' aggregated power consumption time series, the low-frequency parts of the time series are removed to avoid redundancy in the computations. The preprocessed dataset is then fed to the second step, which executes a modified clustering model. Fig. 2.2 shows the overall framework of the proposed phase identification method, which is further elaborated in the following subsections. The two-step preprocessing/clustering approach is then followed by a postprocessing step to check the network's phase mapping. In this regard, the identified phase connectivity is compared to the current network's phase mapping and the potential discrepancies, i.e., wrong phase connectivity information in the current network's phase mapping, are identified and corrected.

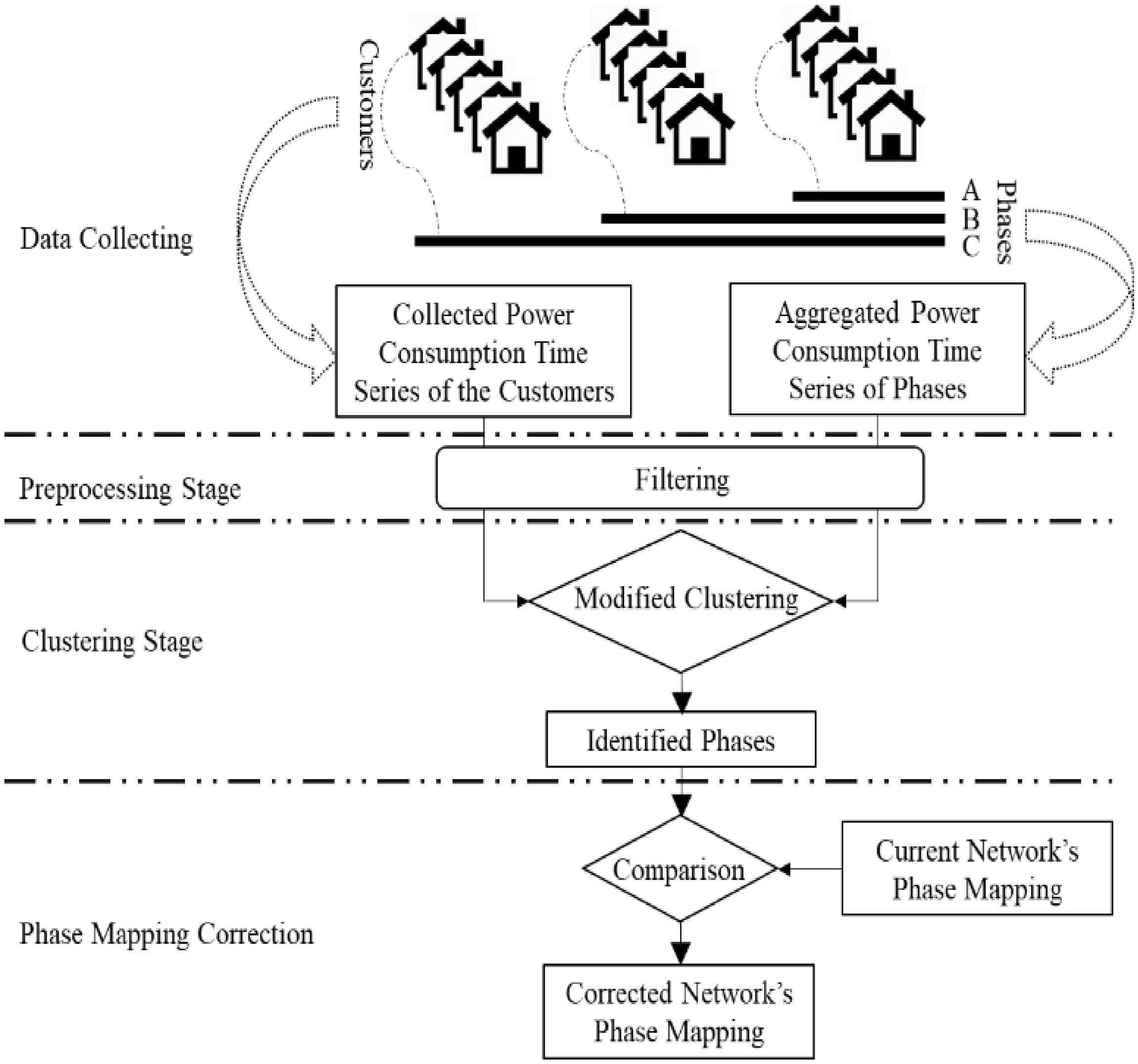

Customers
Phases
A
B
C
Data Collecting
Collected Power Consumption Time Series of the Customers
Aggregated Power Consumption Time Series of Phases
Preprocessing Stage
Filtering
Modified Clustering
Clustering Stage
Identified Phases
Comparison
Current Network's Phase Mapping
Phase Mapping Correction
Corrected Network's Phase Mapping

Fig. 2.2: Proposed phase identification method.

2.3.1 Filtering

In the clustering approach, the input dataset, here called the objects, has a set of potentially unknown features while the label of each object is also not available. The goal is to cluster the data merely based on information found in the data that describes the objects and their relationships. However, some of the features or information contained in the data are redundant, which means these features provide no additional information about the data. Data redundancy leads to data anomalies and corruption and should generally be removed from the data. In other words, important features can help in creating proper clusters, while redundant features may lead to increasing error. In this respect, a set of feature selections are mostly needed to be applied to the data. Feature selection can also reduce the required data size for efficient and accurate clustering [36]. Therefore, a preprocessing step is considered to remove irrelevant and/or redundant features from the data.

The proposed method uses time series of power measurements. Power consumption of the residential customers is always available as the utilities keep the data for billing purposes. Power consumption time series have variations which are resulted from customers' behaviors and lifestyles. Slow-varying components of power consumption follow a similar pattern among residential customers, while fast-varying components are different from customer to customer. In other words, customers are better distinguishable from each other by their high-frequency parts of the power consumption data, i.e., fast-varying components, rather than their low-frequency parts. Using raw power consumption data without preprocessing makes the clustering task to be arduous, as all the customers' power consumption data have similar slow-varying components in their patterns. In contrast, keeping the high-frequency part of the data helps in extracting power consumption patterns. Thus, clustering approach can easily identify each customer's pattern within the aggregated power consumption data of its associated phase. The proposed method uses a high-pass filter to remove low-frequency part of the time series. Fig. 2.3 shows the Bode magnitude and phase plots of the frequency response of a high-pass filter. The cutoff frequency of a filter is a frequency that describes a boundary between a passband and a stopband. In other words, the magnitude of the filter's frequency response at the cutoff frequency is -3 dB of its nominal passband magnitude, where a fall of -3 dB leads to a fall of one-half of the passband power. As it is shown in Fig. 2.3, the magnitude of the filter's frequency response before the cutoff frequency is less than -3 dB, then by applying this filter to the data, the data's frequency band below the cutoff frequency will be practically stopped instead of being passed by the filter which means that part will be eliminated from the data.

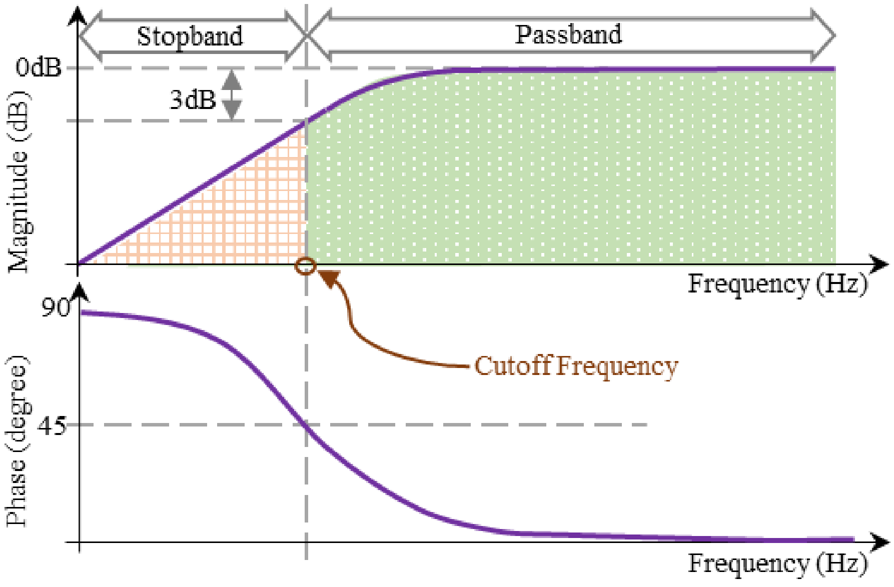


Fig. 2.3: Bode magnitude and phase plots of a high-pass filter.

2.3.2 Modified Clustering Algorithm

Data clustering is an unsupervised and statistical data analysis method for classifying objects with similar features into a homogeneous cluster while objects with dissimilar features are grouped in different clusters. By discovering hidden patterns and relationships associated with the dataset, the clustering algorithm could efficiently classify the objects in proper clusters [37]. Data clustering can be defined as an optimization problem where the objective is to simultaneously maximize the similarity of the objects at the same cluster and minimize the similarity of the objects that belong to different clusters. Similarity can be represented by a proper distance definition, so the optimization can be rewritten as minimizing the intra-cluster distances and simultaneously maximizing the inter-cluster distances as defined in Fig. 2.4.

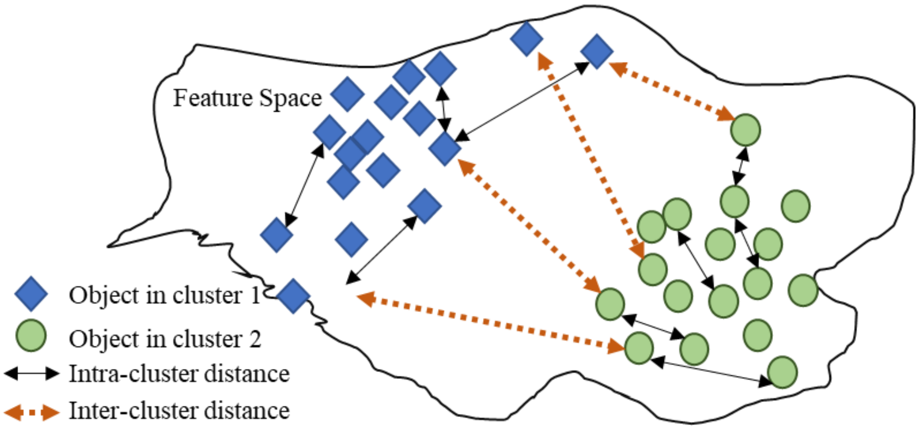


Fig. 2.4: Intra- and inter-cluster distances in feature space for an illustrative clustering example with two clusters.

The traditional K-means clustering algorithm, as a popular data clustering method, divides N objects into K clusters. In this algorithm, initial clusters' centers called centroids are randomly chosen and the similarity between the objects and the centroids is evaluated based on a proper distance formulation such as Euclidean distance. The centroids are accordingly updated by calculating the mean of the previously clustered objects. K-means clustering algorithm can be presented as the optimization problem in (2.1) where x_i is ith object, c_k is the centroid of the kth cluster, and f_D is the proper distance formulation.

$$min \sum_{k=1}^{K} \sum_{j=1}^{N} f_D(x_i, c_k) \tag{2.1}$$

The traditional K-means clustering algorithm is limited to specific applications. However, in this dissertation, a modified clustering algorithm is proposed. Given the fact that the electricity consumption of each residential customer is related to the aggregated electricity consumption of the connected phase at the substation, the correlation concept is utilized to present similarity of the residential customers (2.2), where D_{jk} is the distance between jth residential customer and kth cluster's centroid. In addition, the *Corr* function returns the pairwise linear correlation coefficient between x_j and c_k. In this dissertation, the Pearson type in (2.3) is used, as it leads to higher accuracy when compared to other types of correlation, where σ_{x_j} and σ_{c_k} are the standard deviations of x_j and c_k, respectively.

$$D_{jk} = 1 - Corr(x_j, c_k) \tag{2.2}$$

$$\rho(x_j, c_k) = \frac{cov(x_j, c_k)}{\sigma_{x_j}\sigma_{c_k}} \tag{2.3}$$

The proposed modified clustering algorithm consists of three parts, including initializing centroids, distance calculations, and centroids' updates as demonstrated in Table 2.1.

In the proposed method, instead of randomly initializing the centroids, the aggregated power consumptions of the phases are set as the initial centroids. Then all the residential customers' distances to the three centroids are calculated based on (2.2), and the residential customer with the lowest distance to one of the centroids is selected as the first measurement for phase identification. The minimization in this step is calculated over all clusters and residential customers. On the other hand, compared to other residential customers, the selected residential customer in this step has the least distance to one of the clusters, which means it can be clustered by a high level of confidence. In Step 3, the distances of the selected measurement, i.e.; *j'*th measurement, to the three centroids are evaluated and the minimum distance is found. The selected measurement belongs to the cluster whose centroid is closest to the measurement. As the selected measurement is clustered, it will be eliminated from the data, and the centroids are updated by subtracting the selected measurement from the previous centroids, as in Steps 4 and 5. By repeating these steps for all *N* residential customers, the entire distribution network will be clustered.

Table 2.1: Modified clustering algorithm.

Modified Clustering Algorithm

Inputs:

S_{it}: The aggregated power consumption of phase *i* at the substation,

P_{nt}: The power consumption time series of residential customer *n*,

Parameters:

$t = 1, \dots, T$: Index for time,

i and $k = 1, \dots, K$: Index for cluster, where in this problem *K* is 3.

n, *j*, and $j' = 1, \dots, N$: Index for residential customer/measurement,

N: Total number of residential customers (unknown),

M_{nt}: *n*th measurement,

C_{it}: The centroid of cluster i,

Algorithm:

Step 1) Set S_{it} as the centroids:

$$C_{it} = S_{it}$$

Step 2) Calculate distances between all residential customers and centroids:

$$D_{ji} = f_D(M_{jt}, C_{it}) \forall j, \forall i$$

Step 3) Find the residential customer *j* with the least distance to one of the centroids. This residential customer has the smallest distance to one of the centroids, which means with high possibility, it belongs to the associated cluster. Select the *j*th residential customer's power consumption time series as the *j'*th measurement to be clustered:

$$\min_{j,i}(D_{ji})$$

$$M_{j't} = P_{jt}$$

Step 4) Assign the proper cluster to the *j'*th measurement taking into account that the centroid of the appropriate cluster has the minimum distance to the *j'*th measurement. The *j'*th measurement is clustered in the *k*th cluster, so:

$$\min_{i}(D_{j'i})$$

$$L_{j'} = k$$

Step 5) Remove the *j'*th residential customer from the residential customers' set.

Step 6) Update the centroids:

$$otherwise \quad C_{it} = S_{it}$$

Step 7) Go to step 2 and repeat the loop until all N residential customers are clustered.

Outputs:

L_j: The assigned cluster to the jth residential customer.

2.3.3 Phase Identification in the Presence of Residential Solar PVs

The global environmental concern regarding the use of fossil fuels in electricity generation has motivated many countries to deploy higher levels of renewable energy resources. Among renewable energy resources, solar photovoltaic (PV) is envisioned to be a major player in future power systems and a viable enabler of sustainable power generation. Solar energy is clean, widely available, and relatively low maintenance. Moreover, unlike traditional power generation resources, which are installed in a centralized manner, solar energy resources can be easily deployed as a distributed generation resource [38]-[42]. Solar energy resources have attracted consumers who are willing to make up part of their electricity consumption or even economically benefit from a local power generation [43], [44]. The dropping cost of solar technology and the state and government incentives have made the path for rapid growth of solar generation. More than 7 GW of solar PV was installed in the U.S. in 2016, where residential PV with over 2 GW represented the biggest segment [45], [46]. All in all, solar generation is making fast inroads in power systems [47]-[49].

Solar photovoltaic (PV) is facing a significant cost reduction due to the technical advances in its technology combined with increased market demand. As a result, PV penetration is growing across the world [47], [50]. The positive public support for installing PV units is one of the motivations to introduce supportive policies for solar energy in many regions of the U.S. This increasing demand further leads to a decline in the costs associated with the PV installation. Fig. 2.5 shows the residential solar PV installations and forecasts for 2015-2024.

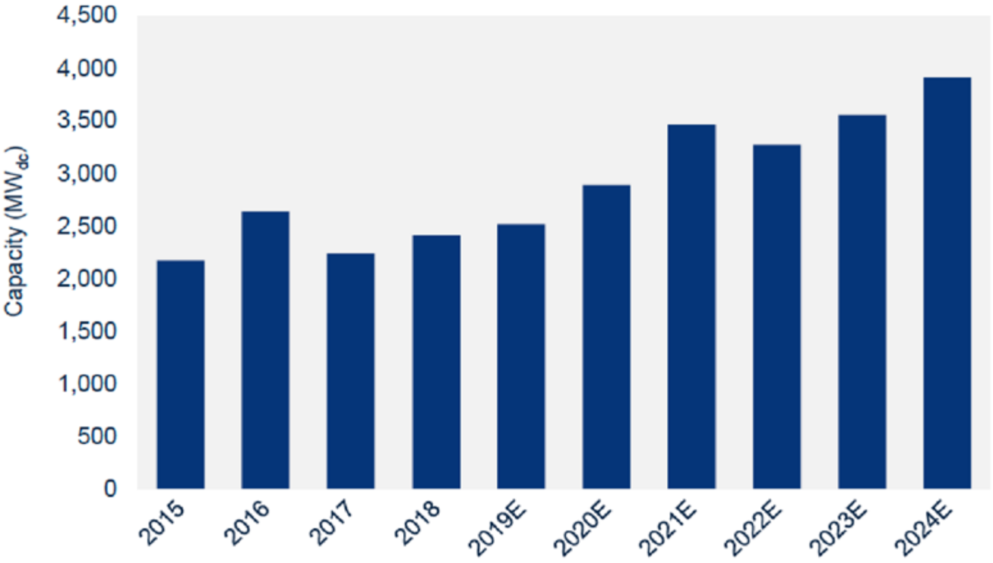


Fig. 2.5: Residential solar PV installations and forecasts for 2015-2024 [51].

Growing penetration of non-dispatchable energy resources, i.e., PV units, causes technical and operational challenges for the utilities and distribution grids [43], [47]. Integrating solar systems into the current distribution grids may add additional difficulties regarding the phase identification task. However, the proposed phase identification method is designed to handle residential solar PVs in the distribution network.

2.4 Numerical Studies

The proposed method is applied to both small and relatively large networks under scenarios of complete and incomplete data. In Case 1, by considering complete data to evaluate the scalability of the proposed method, the proposed method is tested on the IEEE 25-bus test system and a 450-bus distribution network with unknown topology. The IEEE 123-bus distribution test system is further used in Case 2 to examine the ability of the proposed method in case of incomplete data. Similarly, Case 3 uses the IEEE 123-bus distribution test system to evaluate the effectiveness of the proposed method in the presence of residential solar PVs. The power consumption data is borrowed from [52], which is publicly available. However, based on the network topology for each case (25-bus, 123-bus, or 450-bus), the aggregated power consumption for each phase has been manually calculated and by using power flow simulation, the effect of loss and the network configuration has been applied to the data. Calculations were done in MATLAB [53] and GAMS [54]. Solar power data used in Case 3 is a modified version of data publicly available in [55]. In this regard, a set of scaling and preprocessing has been applied to the solar power data based on the peak load of each residential customer. Preprocessed solar power data has been then added to the selected residential customer's power consumption as well as to the aggregated power consumption of the corresponding phase. Moreover, the effect of loss and the network configuration has been applied to the final data, as previously explained.

Case 1: Phase identification based on complete data

The proposed method is tested on the IEEE 25-bus distribution test system as well as a 450-bus distribution network with an unknown topology. Power consumption of the residential customers is collected at 30-min time intervals, while the aggregated power consumption of the three phases at the substation is also available for the same time intervals. A high-pass filter is applied to the recorded time series. In this dissertation, the high-pass filter is designed in MATLAB as a 50th-order high-pass window-based FIR (finite impulse response) filter, as shown in Fig. 2.6. As the sampling frequency (sample/sec) of the power consumption time series is low, the normalized cutoff frequency is considered to be 0.35 (× π rad/sample), to remove the lower frequency band while avoiding losing the useful portions of the data. It should be noted that by using the Nyquist frequency (half of the sampling frequency), the normalized frequency can be obtained.

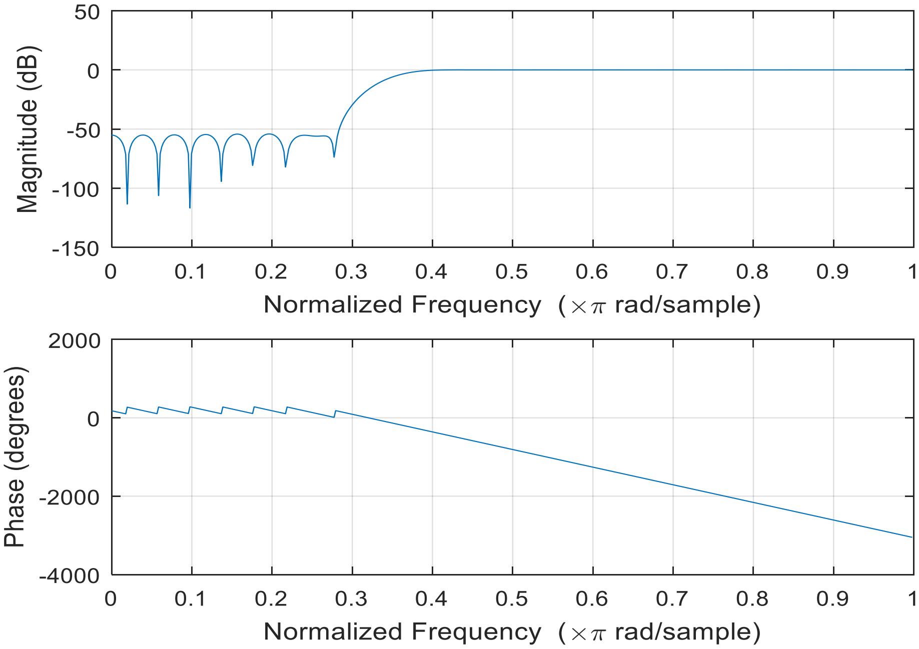


Fig. 2.6: Bode plot of the used high-pass filter.

2.4.1 IEEE 25-Bus Distribution Test System

Fig. 2.7 shows the IEEE 25-bus distribution test system. All 25 buses in this system are unbalanced three-phase, and in this case, it is assumed that all three phases are balanced in terms of the number of connected residential customers. In this respect, among the total number of residential customers, i.e., 75 residential customers, 25 random customers are connected to each phase. Based on the available data, the average loads over the 30-day time period for phases A, B, and C are measured as 11.30 MW, 14.62 MW, and 13.11 MW, respectively.

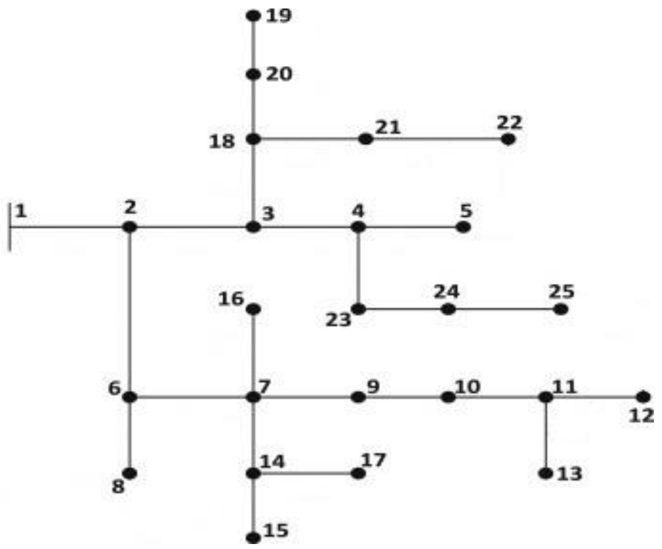


Fig. 2.7: IEEE 25-bus distribution test system.

Residential customers' power consumptions are collected in 30-min time intervals through smart meters installed at each residential customer. The aggregated power consumptions at all three phases at the substation are also available. At the first step, ineffective and redundant parts of the power consumption time series are removed by filtering the low-frequency part of the time series for each residential customer, as well as for each phase. Fig. 2.8 shows the power consumption of one of the residential customers in the time domain and its single-sided spectrum in the frequency domain before and after filtering. As shown in this figure, the low-frequency portion of the data is eliminated to remove redundant parts.

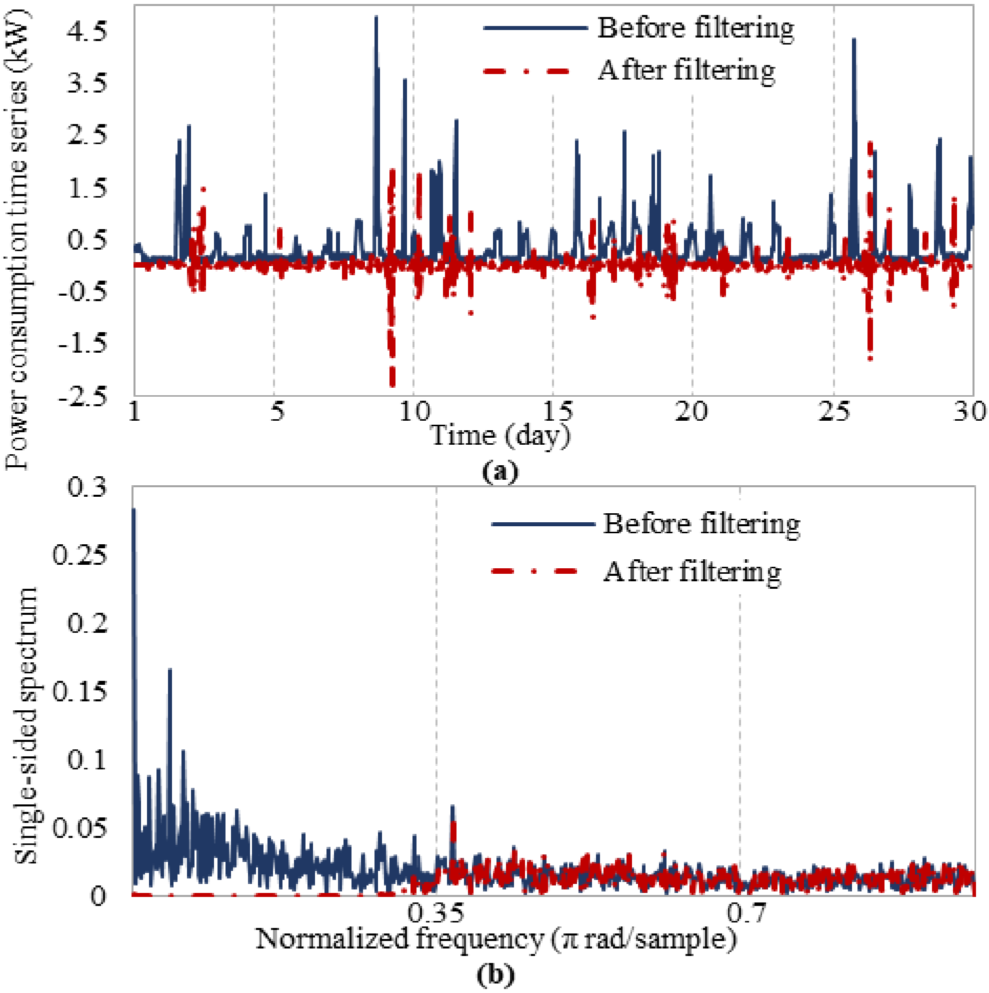


Fig. 2.8: (a) Power consumption of a residential customer in the time domain and (b) its single-sided spectrum in the frequency domain before and after filtering.

The preprocessed power consumption time series of the residential customers and three phases are then applied to the modified clustering algorithm. The preprocessed power consumption time series of each residential customer is strongly correlated to the preprocessed aggregated power consumption time series of the connected phase at the substation. In other words, the distance of a residential customer's power consumption to the connected phase is smaller than its distance to other phases.

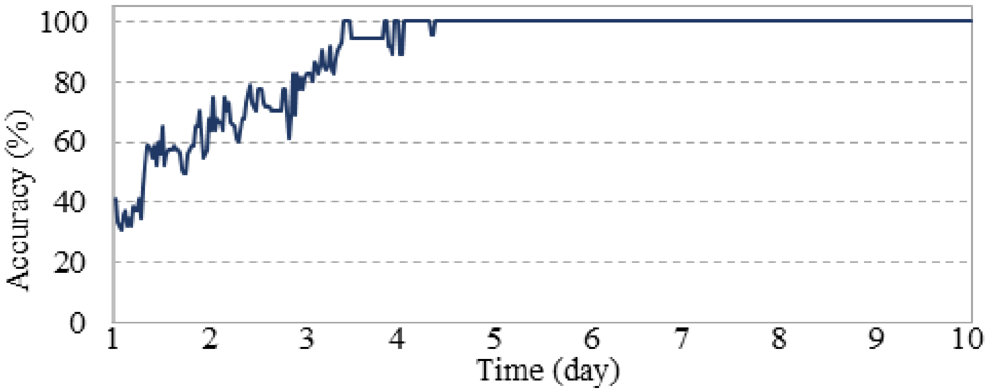


Fig. 2.9: Phase identification method's accuracy by increasing the number of days of available data for the IEEE 25-bus distribution test system.

Fig. 2.9 shows the solution accuracy by increasing the number of days of available data for the IEEE 25-bus distribution test system. As shown, the phase connectivity of all the residential customers in the IEEE 25-bus distribution test system is identifiable by having at least 210 samples, which is equal to 4 days and 9 hours of 30-min power consumption measurements. However, the length of the required data is related to the network size, and by increasing the size of the network, the length of required data for phase identification needs to be adequately large. On the other hand, by increasing the resolution, i.e., reducing the recording time interval, the length of required data will be reduced.

2.4.2 450-Bus Distribution Network

In this case, it is assumed that the topology of the network is unknown. The network includes 450 residential customers where 200, 100, and 150 residential customers are connected to phases A, B, and C, respectively. It is assumed that the network is large and extremely unbalanced where the average loads over the 30-day time period for phases A, B, and C are 183.99 MW, 120.72 MW, and 84.77 MW, respectively. To demonstrate the effectiveness of the proposed preprocessing step, the proposed method is applied to this system in two scenarios of (i) ignoring and (ii) considering the preprocessing step. The available data is assumed to change from 1 day to 30 days by steps of 1 hours. Fig. 2.10 shows the accuracy of the proposed method in these two scenarios.

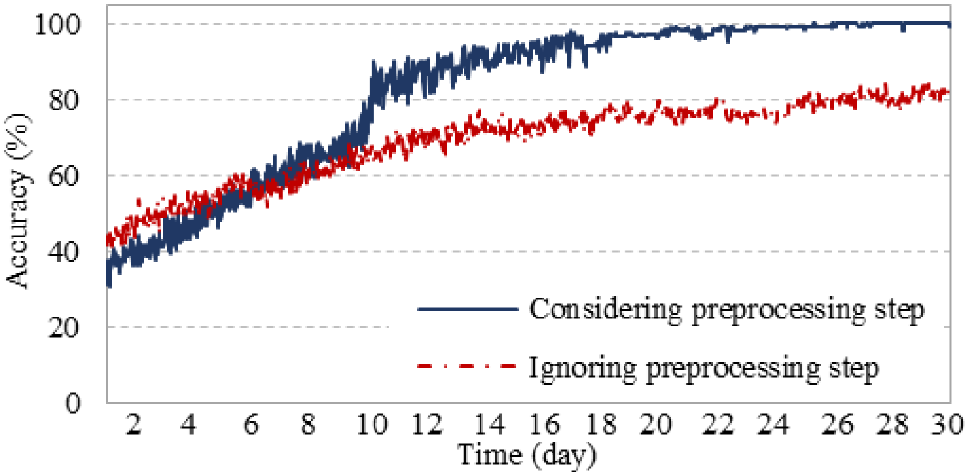


Fig. 2.10: Phase identification method's accuracy for 450-bus distribution network by ignoring and considering preprocessing step.

As shown in Fig. 2.10, the preprocessing step significantly improves the phase identification results. When more than 24 days of power consumption data are available for the studied network, applying preprocessing step to the data leads to phase identification with the accuracy of 100%, while without preprocessing step, it will remain below 85%.

It should be mentioned that the computation times for identifying the phase connectivity of residential customers by the proposed method for the IEEE 25-bus distribution test system and the 450-bus distribution network are 0.25 s and 4.55 s, respectively. In addition, the results in Case 1 show that by increasing the network size or decreasing the resolution, the amount of required data would considerably increase, however, the proposed method is scalable and capable of finding the phase connectivity of customers for much larger systems, whether balanced or not, in a relatively short amount of time.

Case 2: Phase identification based on incomplete data

The power consumption data of a residential customer recorded by a smart meter can be lost for various reasons. For example, the communication for a group of residential customers in a feeder could be disconnected. As a result, the power consumption data of mentioned residential customers could be lost for specific time intervals. In case of data loss, the power consumption at known time intervals is replaced with zero. However, the aggregated power consumption data recorded for each phase at the substation is accurate and complete. While the proposed method is able to identify each customer's phase connectivity individually without having the other residential customers' information, the impacts of incomplete data can be investigated for a single residential customer as well as a group of residential customers. In this respect, by considering the IEEE 123-bus test system, two kinds of incompleteness are considered, i.e., random and consecutive incompleteness.

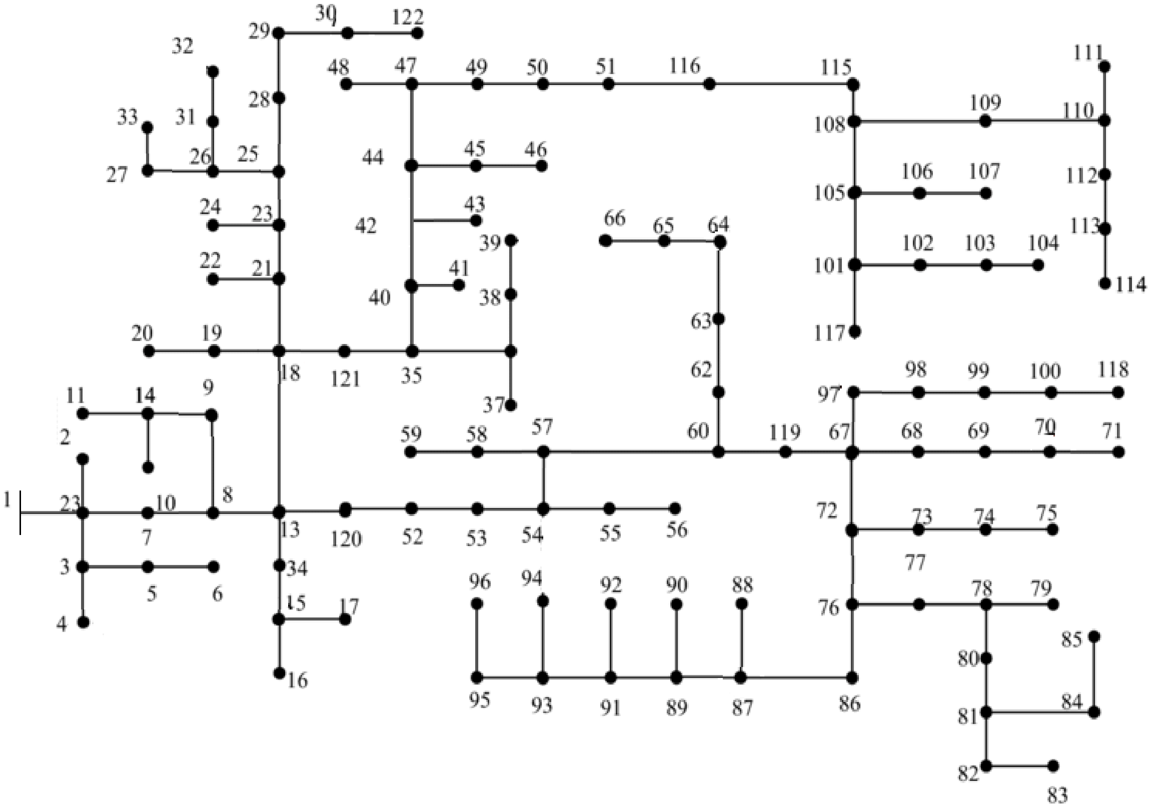


Fig. 2.11: IEEE 123-bus distribution test system.

Fig. 2.11 shows the schematic diagram of the IEEE 123-bus distribution test system. Among 123 buses in this system, 85 buses are load buses in which 5 are unbalanced three-phase and the rest are single-phase. It is assumed that 30, 25, and 40 of the residential customers are connected to phases A, B, and C, respectively. Based on the available data, the average loads over the 30-day time period for phases A, B, and C are 17.89 MW, 17.39 MW, and 12.39 MW, respectively. In this case, by accessing at least 6 days of power consumption data of residential customers without any incompleteness for the test system, the proposed method is able to identify phase connectivity of residential customers by the accuracy of 100%.

2.4.3 Random Incompleteness

In this case, it is assumed that the power consumptions of a set of residential customers are not recorded by the associated smart meter in random time intervals. Randomness is modeled by a uniformly distributed pseudorandom integer producer. Sensitivity analysis with respect to the percentage of incomplete time intervals and the percentage of residential customers with incomplete data is performed for the IEEE 123-bus test system. It is considered that 14 days of 30-min power consumption data for residential customers as well as for each phase at the substation are available. The percentage of incompleteness changes from 0% to 20% by steps of 2%, while the number of residential customers with incomplete data increases from 10% to 100% by steps of 10%. Fig. 2.12 shows the real power consumption time series of a residential customer for 14 days as well as recorded data with 20% random incompleteness. In this case, data at 134 random time intervals out of total 672 time intervals are not sent because of the lack of communication during those time intervals.

The results of sensitivity analysis are summarized in Table 2.2. As it is expected, in the case of random incompleteness, when the percentage of random incompleteness in each residential customer increases from 0% to 20%, accuracy is slightly decreased. On the other hand, when the number of residential customers with incomplete data increases from 10% (10 out of 95 residential customers) to 100% (all 95 residential customers), the accuracy of the proposed method varies between 100% and 80%. However, the proposed method is capable of identifying residential customers' phases with acceptable accuracy.

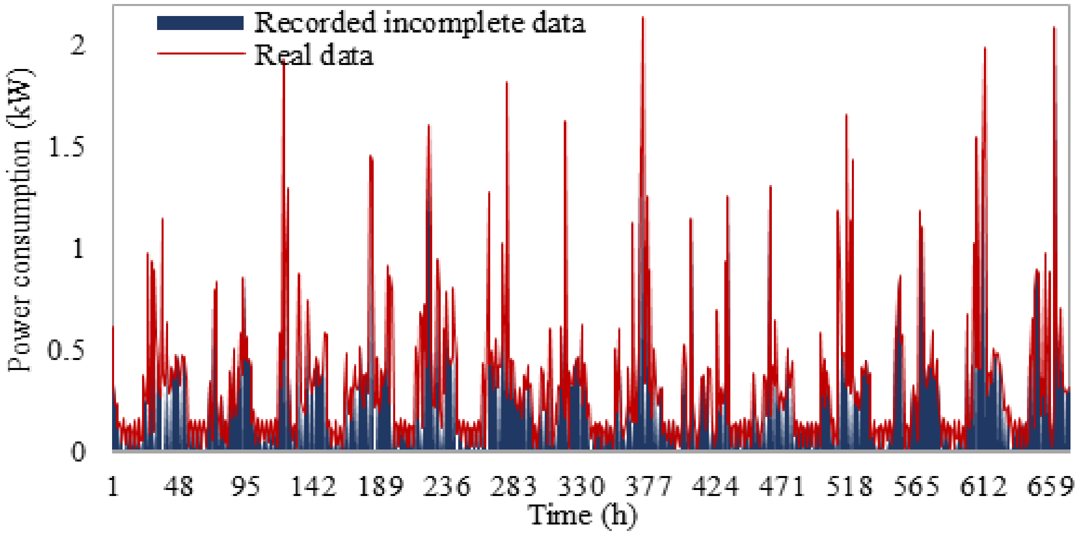


Fig. 2.12: Real power consumption time series of a residential customer for 14 days and recorded data with 20% random incompleteness.

Table 2.2: Accuracy of proposed method (%) with respect to the number of residential customers with incomplete data and percentage of random incompleteness.

		Number of residential customers with incomplete data (%)									
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Random incompleteness for each residential customer (%)	**0**	100	100	100	100	100	100	100	100	100	100
	2	100	100	100	100	100	98.95	100	98.95	97.89	100
	4	100	100	100	100	100	97.89	95.79	98.95	96.84	94.74
	6	100	97.89	100	97.89	96.84	93.68	95.79	94.74	94.74	93.68
	8	98.95	97.89	98.95	100	93.68	95.79	95.79	97.89	90.53	94.74
	10	100	100	96.84	96.84	93.68	93.68	90.53	87.37	88.42	96.84
	12	98.95	95.79	97.89	96.84	96.84	90.53	90.53	90.53	93.68	85.26
	14	98.95	97.89	97.89	96.84	95.79	92.63	91.58	94.74	86.32	88.42
	16	98.95	96.84	93.68	92.63	89.47	92.63	88.42	90.53	92.63	89.47
	18	95.79	97.89	95.79	97.89	89.47	89.47	91.58	88.42	85.26	84.21
	20	100	95.79	91.58	97.89	93.68	90.53	90.53	91.58	83.16	83.16

2.4.4 Consecutive Incompleteness

In this case, consecutive incompleteness in recorded power consumption of a set of residential customers is considered. Similar to the previous scenario, sensitivity analysis with respect to the percentage of incompleteness is performed for the IEEE 123-bus test system. The percentage of incompleteness changes from 0% to 20% by steps of 2%, while the number of residential customers with incomplete data increases from 10% to 100%. Fig. 2.13 shows the real power consumption time series of a residential customer for 14 days as well as recorded data with 20% random incompleteness.

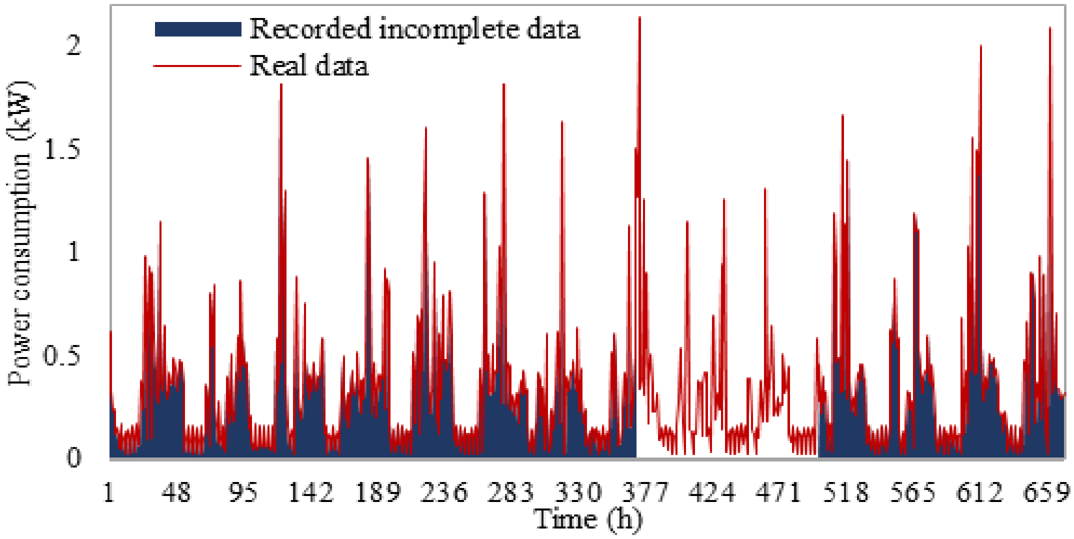


Fig. 2.13: Real power consumption time series of a residential customer for 14 days and recorded data with 20% consecutive incompleteness.

The results of sensitivity analysis are summarized in Table 2.3. Compared to the previous scenario, in this scenario, incompleteness is limited to a consecutive time interval and the residential customers' power consumption for other time intervals is not extremely affected by the missing data. As a result, the proposed method is able to effectively identify residential customers' phases almost for all possible scenarios with an accuracy of more than 94.74%. The worst-case scenario occurs when all residential customers are experiencing 20% of consecutive incompleteness, however, the proposed method is still able to correctly identify phase connectivity of 90 out of 95 residential customers.

Table 2.3: Accuracy of proposed method (%) with respect to the number of residential customers with incomplete data and percentage of consecutive incompleteness.

		Number of residential customers with incomplete data (%)									
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Consecutive incompleteness for each residential customer (%)	**0**	100	100	100	100	100	100	100	100	100	100
	2	100	100	100	100	100	100	98.95	97.89	100	100
	4	100	100	100	100	100	100	98.95	100	100	100
	6	100	100	100	100	100	98.95	98.95	100	98.95	100
	8	100	100	97.89	98.95	97.89	98.95	97.89	98.95	98.95	97.89
	10	100	100	100	98.95	98.95	97.89	100	97.89	100	98.95
	12	100	100	98.95	100	98.95	98.95	100	95.79	97.89	95.79
	14	100	100	100	98.95	98.95	98.95	100	97.89	97.89	96.84
	16	100	100	98.95	98.95	98.95	97.89	94.74	97.89	97.89	96.84
	18	98.95	100	100	97.89	97.89	97.89	97.89	97.89	95.79	96.84
	20	100	98.95	100	97.89	97.89	95.79	97.89	95.79	98.95	94.74

To get a step further, random and consecutive incompleteness are applied to a larger portion of the power consumption data of residential customers. In this regard, the sensitivity analysis is studied with respect to the percentage of incompleteness changes from 0% to 90% by steps of 2%, while the number of residential customers with incomplete data increases from 10% to 100%. Fig. 2.14 and Fig. 2.15 summarize the results for consecutive and random incompleteness, respectively.

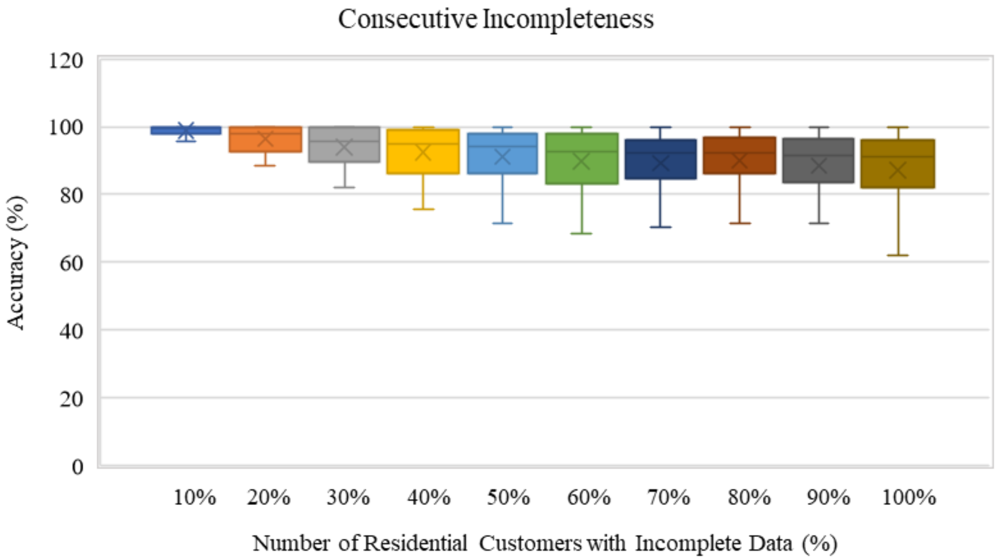


Fig. 2.14: The range of phase identification method's accuracy with respect to the number of residential customers with incomplete data and percentage of consecutive incompleteness.

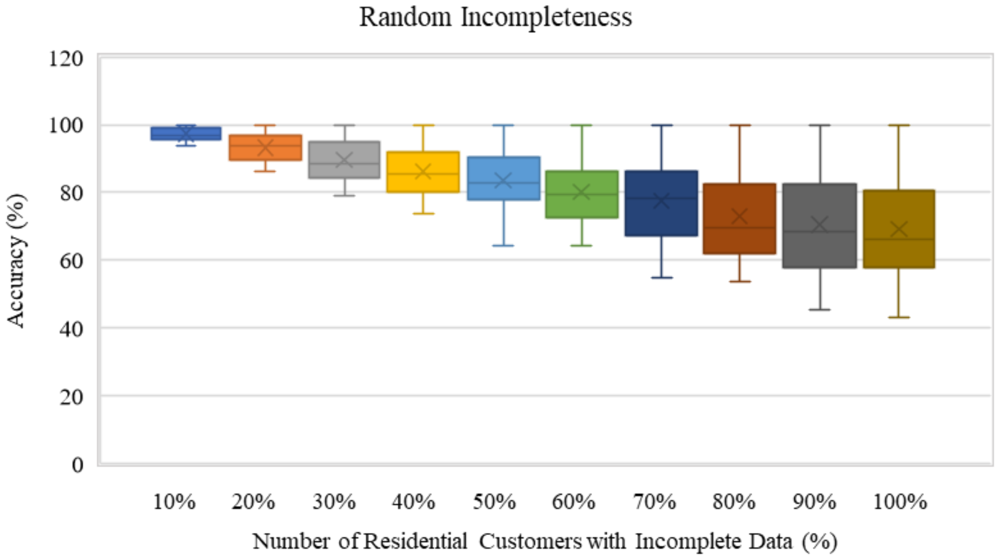


Fig. 2.15: The range of phase identification method's accuracy with respect to the number of residential customers with incomplete data and percentage of random incompleteness.

As Fig. 2.14 shows, by increasing the number of residential customers with consecutive incomplete data, the accuracy of the proposed slightly drops, however, even in the worst-case scenario, when all residential customers are experiencing incompleteness of 10% up to roughly 90%, the proposed method is capable of identifying the phase connectivity. This is the case for random incompleteness shown in Fig. 2.15. However, as previously mentioned, as incompleteness in consecutive type is limited to a consecutive time interval and the residential customers' power consumption for other time intervals is not extremely affected by the missing data, the proposed method has a better performance compared to random incompleteness. This can be seen by comparing the results in the last two figures, where by increasing the number of residential customers with random incomplete data in each scenario, the accuracy of the proposed method decreases slightly faster than in case of consecutive incomplete data.

Case 3: Phase identification considering residential solar PVs

Integrating solar systems into the current distribution grids may add additional difficulties regarding the phase identification task. The proposed phase identification method in the presence of residential solar PVs is evaluated in this dissertation. In this regard, the IEEE 123-bus test system is used while considering PV installation in selected buses.

2.4.5 Distribution Test System with Residential Solar PVs

In this case, it is assumed that a set of random residential customers are equipped with solar PVs. Randomness is modeled by a uniformly distributed pseudorandom integer producer. Sensitivity analysis with respect to the length of available data and the percentage of residential customers with solar PV installation is performed for the IEEE 123-bus test system. It is considered that 30-min power consumption data for residential customers as well as for each phase at the substation are available. The length of available data is increased from 1 day to 10 days by steps of 1 day, while the number of residential customers with solar PVs increases from 0% to 50% by steps of 10%.

The results of sensitivity analysis are summarized in Table 2.4. As shown, by having at least 7 days of data, the accuracy of the proposed method reaches its max of 100%. As it is expected, as the PV installation is a behind-the-meter energy resource, the net load metered by the smart meter remains unchanged. As a result, the proposed method is efficiently capable of identifying the phase connectivity of the customers in this case.

Table 2.4: Accuracy of the proposed method (%) with respect to the number of residential customers with solar PVs and the number of available data.

		Number of available data (day)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Number of residential customers with solar PVs (%)	**0**	47.37	65.26	73.68	94.74	98.95	100	100	100	100	100
	10	46.32	74.74	74.74	97.89	100	100	100	100	100	100
	20	38.95	66.32	77.89	92.63	100	100	100	100	100	100
	30	48.42	56.84	75.79	93.68	98.95	100	100	100	100	100
	40	41.05	52.63	75.79	93.68	97.89	100	100	100	100	100
	50	45.26	64.21	76.84	94.74	97.89	98.95	100	100	100	100

2.4.6 Discussion

Several interesting points can be derived regarding the proposed method based on the studied cases:

- As it can be seen in Tables 2.2 and 2.3, in each column, by increasing the percentage of incompleteness, the accuracy does not monotonically decrease. The reason is that to evaluate the capability of the proposed method in case of incomplete data, in each step the selected residential customers carrying incomplete data are chosen randomly to make sure most of the possible combinations are tested. In other words, in each cell of each column in Tables 2.2 and 2.3, the residential customers with incomplete data are selected randomly and the samples lost in each power consumption time series are also chosen randomly. This way, the accuracy of each step of the sensitivity analysis is independent of the accuracy of previous steps.
- As it is expected, increasing the network's size leads to an increase in the length of required data for phase identification.
- Increasing the percentage of incompleteness in the recorded power consumption of residential customers leads to a decrease in the proposed method's accuracy, which can be compensated by increasing the length of the available data.
- Increasing the number of residential customers with incomplete data also leads to a decrease in the proposed method's accuracy, which again can be offset by increasing the length of the available data.

- Power consumption is considered to be recorded in a timely manner at 30-min time intervals. Recording data in shorter time intervals could decrease the length of required data while expediting the clustering.

- There is no assumption that the dataset should be complete. However, by accessing to *a priori* information regarding the residential customers with incomplete data and modifying the proposed method, the results could be improved. This can be achieved through dividing the proposed method to two steps, one for the residential customers with complete data and the next for those with incomplete data.

- As it can be seen in Tables 2.4, the proposed method is efficiently capable of identifying the phase connectivity of the residential customers in the presence of solar PVs. Smart meters collect and report the net load consumed by the residential customer. On the other hand, the meters installed at the substation also record the net load by all the customers connected to each phase. The proposed phase identification method uses the net load recorded by the smart meters, so the method is capable of identifying the phase connectivity of the residential customers in the presence of PV or any other behind-the-meter energy resource such as batteries or EVs.

2.5 Conclusion

An innovative data-based phase identification method was proposed in this dissertation. The proposed method consisted of two steps of preprocessing and clustering. In the preprocessing step, redundant and useless parts of the power consumption time series were removed by using a high-pass filter. The preprocessed data were then applied to the clustering algorithm. An efficient clustering algorithm was also developed starting from the aggregated power consumption of the three phases at the substation as the initial centroids, where the residential customers were accordingly assigned to the corresponding clusters. In addition, by applying the postprocessing step, the proposed method identified and corrected the wrong phase connectivity information in the current network's phase mapping available at the electric utility.

The effectiveness of the proposed method to identify the residential customers' phases in case of incomplete data was evaluated by considering two possible scenarios of incomplete data, i.e., random and consecutive. In addition, the performance of the proposed method in the presence of residential solar PVs was analyzed. Numerical results showed that the proposed method could identify the phase connectivity of the residential customers individually and accurately. The proposed phase identification method can be utilized by electric utilities to improve observability and controllability of the distribution grid as an important part of topology identification for different economic and operational reasons, including but not limited to rebalancing, phase mapping connection, loss reduction, operational improvements, and hosting capacity calculations.

3 Chapter Three: Distribution Outage Identification

3.1 Introduction

Extreme weather events can cause significant damage to electric power grid infrastructure and lead to widespread power outages. The frequency and the intensity of these events are continuously increasing as a direct result of climate change [12]. From 2002 to 2015, extreme weather events have caused more than 87% of major power outages, involving 50,000 or more customers, in the U.S. [13]. Fig. 3.1 shows the most common causes of the outages in the US in 2020, where more than 43% of them were weather- or natural disaster-related.

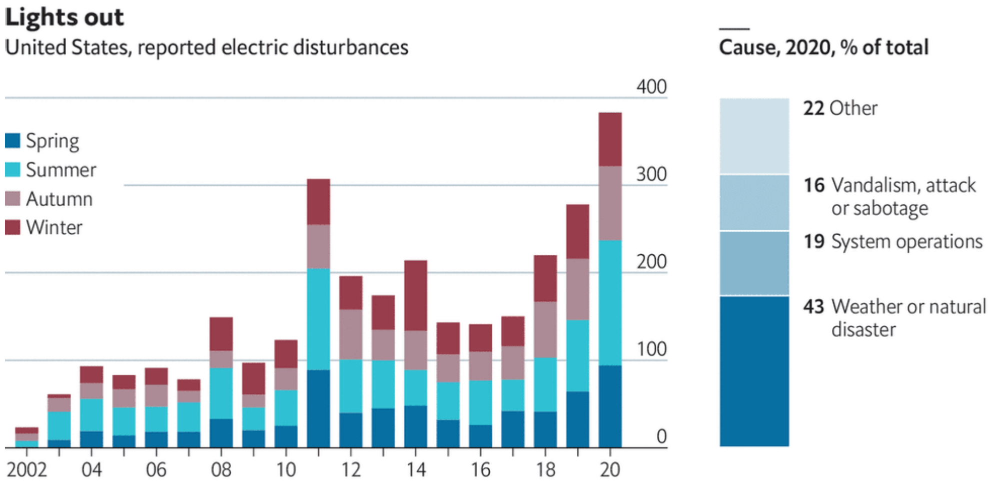


Fig. 3.1: Common cause of electricity outages in the US in 2020 [56].

The identification of grid components that are damaged is the first step to recovering from extreme weather-related power outages. These components can be easily identified in the generation and transmission levels; however, this is not the case for distribution level components. Traditionally, the distribution grid has not been fully observable to grid operators, causing outage location identification a challenging task. This task is currently performed by investigating the feeder configuration map and the protection design manual to identify the overall outage locations [14], i.e., an expert-experience-based method. Although the expert-experience-based method may be able to achieve highly accurate solutions, it is proved to be laborious, costly, and time-consuming [15], [16].

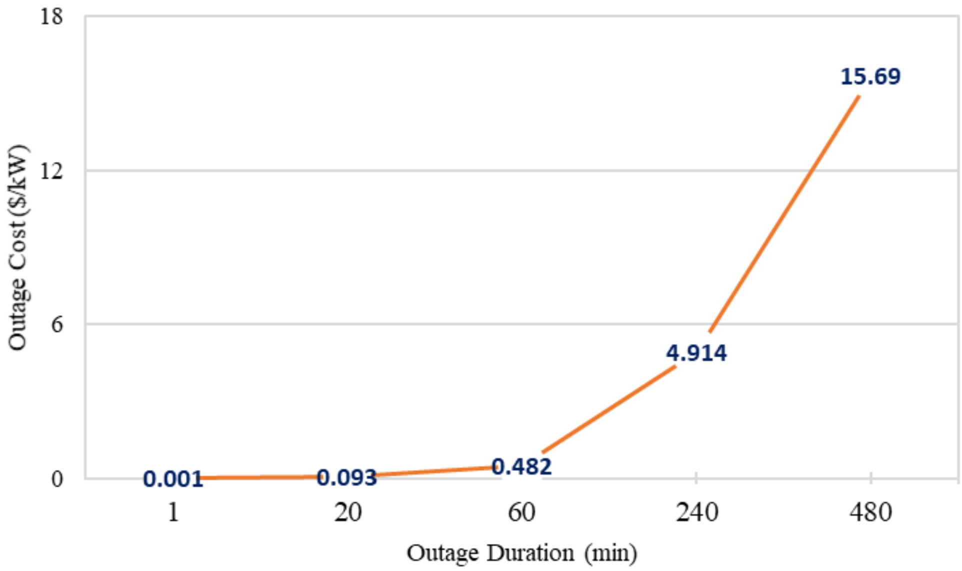


Fig. 3.2: Outage cost vs. outage duration.

Time is of particular essence in this case. As the outage duration increases, the associated outage cost will increase almost exponentially, as shown in Fig. 3.2. An automated method with a short computation time would be significantly useful in this case, especially for larger networks.

Consider a distribution network in which one or more lines are disconnected due to an extreme weather event. Because of the radial structure of the network, multiple customers will experience a power outage. Assuming each customer is equipped with a smart meter, a signal (commonly known as the "last gasp", hinting that the meter will go offline after this signal) will be instantaneously sent to the electric utility company. Once these last gasp signals are received, the utility company should figure out the exact location(s) of the damaged line(s) and accordingly proceed to restore the distribution grid in the shortest possible time. The key challenge here is that depending on the distribution grid topology, discovering the exact locations of damaged lines can be a complicated and time-consuming task.

An important issue regarding weather-related outages in distribution grids is that several lines may be damaged simultaneously in various locations, where some noteworthy methods, such as in [57] and [58], are proposed for identifying multiple simultaneous outages. Fig. 3.3 depicts an illustrative example in a 13-bus radial distribution grid. As illustrated, customers on buses 8 to 13 are experiencing power outages which are automatically reported to the utility company by smart meters. The challenge here is that a power outage can result from damage to any of the upstream distribution lines. For example, for a power outage in bus 13, any of the seven upstream lines can be damaged, while for an outage in bus 8 this number is eight. This example advocates that the damaged lines can be mapped into multiple target labels. The idea in this dissertation is to figure out the exact lines that cause the observed power outages [7].

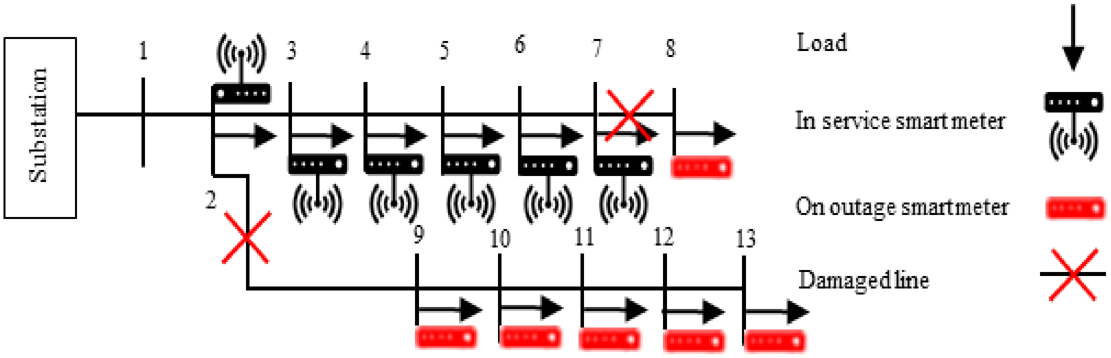


Fig. 3.3: Example of simultaneous power outages in a radial network.

3.2 Proposed Solution

Considering the potentially large amount of data that will be involved in this process and the frequency of these events, utility companies could reap the benefits of a machine learning-based data mining method in locating the damaged lines. Fig. 3.4 depicts the proposed solution, which consists of four steps; data synthesis, training, testing, and evaluation. Through data synthesis, the historical network topology and the outage status of smart meters are used to label the actually damaged lines. By leveraging a large portion of the synthesized data, the second step reaps the benefit of several binary SVMs to train a multi-label classifier and generate the associated SVM parameters for the test step. The rest of the synthesized data, along with the generated SVM parameters, are employed in the last step to evaluate the effectiveness of the proposed method.

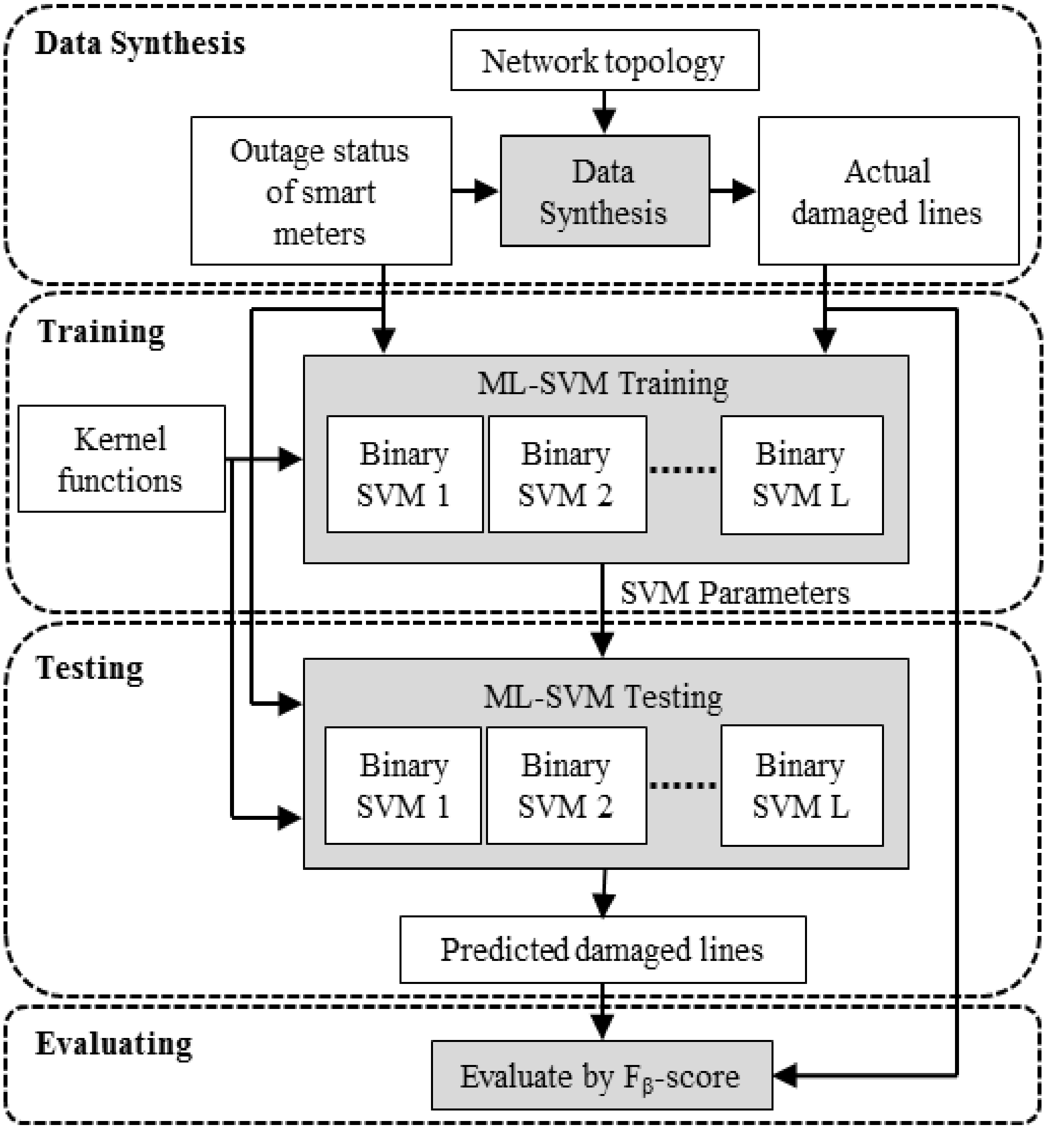


Fig. 3.4: The input and output vector data for the ML-SVM classifier.

3.2.1. Multi-Label SVM Classifier

SVM, which has been widely used in various areas of research [59], was originally designed as a binary classifier. However, a set of independent binary SVMs can be employed to perform multi-label classification [59]. The binary SVM classification is an optimization problem based on finding a hyperplane. The optimal hyperplane minimizes the distance among the training samples which belong to the same class while maximizing the distance among the samples of different classes. As shown in Fig. 3.5, the separating distance between two classes is called a margin, and the closest samples to the hyperplane are called the support vectors [60].

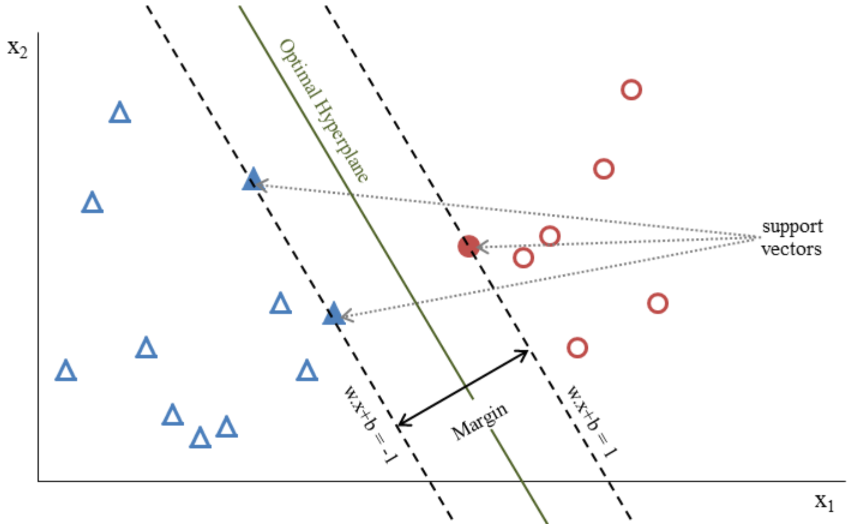


Fig. 3.5: The binary SVM classifier.

The hyperplane can be defined as (3.1):

$$w.x + b = 0 \tag{3.1}$$

where ***w*** and *b* are the SVM model parameters, and ***x*** is the input training data.

Consider a multi-label training set $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, where $\boldsymbol{x}_i$ is a *B*-dimensional input vector representing the outage status of the smart meters. *B* is the number of smart meters, and $x_{id} = 1$ means *d*-th smart meter is in service, while $x_{id} = 0$ indicates this smart meter is reporting a power outage. On the other hand, $\boldsymbol{y}_i$ is the status of the lines as a label vector. Each $\boldsymbol{y}_i$ has *L* arrays which correspond to lines. If $y_{il} = 1$, line *l* is on outage; otherwise, if

y_{il} = -1, this line is in service. By defining ξ_{il} as a nonnegative slack variable, and c as a tradeoff parameter, the ML-SVM classifier can be written as an optimization for each class (3.2):

$$\min \frac{1}{2} \|w_l\|^2 + c \sum_{i=1}^{N} \xi_{il} \tag{3.2}$$

$$subject\ to \quad y_{il}(w_l^T x_i + b_l) \geq 1 - \xi_{il}, \quad l = 1, \dots, L$$

N and L are the number of training data and the number of classes, respectively [60]. The optimization problem returns a binary SVM model for every individual class as in (3.3).

$$f_l(x_i) = w_l^T x_i + b_l \tag{3.3}$$

3.2.2. Nonlinear Classification

Although the original SVM algorithm is a linear classifier, by applying a kernel function to the algorithm, the nonlinear SVM classifier can be achieved [61]. The mapping function $\varphi(\boldsymbol{x}_i)$ transforms the input vector to a higher-dimension space. Fig. 3.6 Shows the transformation from a linear to nonlinear input space.

The nonlinear SVM classifier maximizes the margin by finding the hyperplane in the transformed feature space. The kernel function, i.e., $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a function of $\varphi(\boldsymbol{x}_i)$ as shown in (3.4).

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}_j) \tag{3.4}$$

There are several kernel functions used in SVM, including but not limited to linear [61], [62], polynomial [63], and radial basis function (RBF) [64], [65], which are represented as in (3.5)-(3.7):

$$K^{Linear}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i{}^T \boldsymbol{x}_j \tag{3.5}$$

$$K^{Polynomial}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\gamma_p \boldsymbol{x}_i{}^T \boldsymbol{x}_j + r\right)^{d'} \tag{3.6}$$

$$K^{RBF}(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\left(-\gamma_r \|x_i - x_j\|^2\right)} \tag{3.7}$$

where both γ_p and γ_r are positive parameters, r is a nonnegative parameter, and d' is the degree of the polynomial kernel [60].

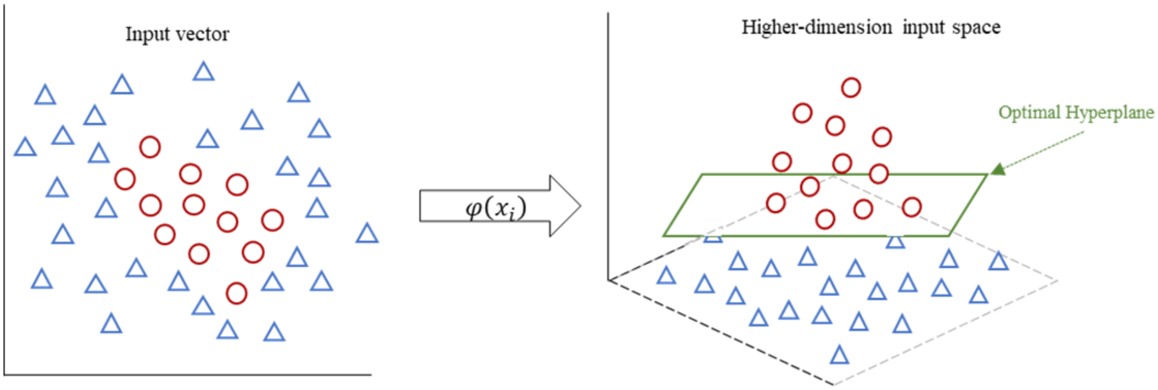


Fig. 3.6: Kernel transformation.

3.2.3. Model Evaluation

A confusion matrix is used to evaluate the performance of the proposed classifier. The predicted label for each sample could acquire four possible states, including True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP), as shown in Table 3.1. The diagonal components demonstrate the number of samples which are correctly classified, while the off-diagonal components represent errors. The FN component is referred to as error type I (E_I), and it illustrates the number of damaged lines which are wrongly classified as in service. The FP component is called error type II (E_{II}), denoting the number of in-service lines which are wrongly categorized as on outage.

Table 3.1: Confusion matrix

		Predicted Class	
		-1 (In service)	1 (On outage)
Actual Class	-1 (In service)	TN	FP
	1 (On outage)	FN	TP

It is interesting to note that based on the problem definition, these errors have different meanings:

- E_I means that the time of outage in some locations will be longer, as no repair personnel will be sent out to those locations
- E_{II} translates into futile restoration efforts, as some repair personnel will be sent out to incorrect locations.

Various measures can be defined to evaluate the classifier performance from the confusion matrix, including F_β-score as adopted in this dissertation (3.8):

$$F_\beta - score = \frac{TP}{TP+\left(\frac{\beta^2}{1+\beta^2}\right)E_I+\left(\frac{1}{1+\beta^2}\right)E_{II}} 0 \leq \beta \leq +\infty \tag{3.8}$$

As formulated in (3.8), when β < 1, the effect of E_{II} will be greater than E_I, for β = 1, the impact of these two errors is equal to each other, and when β > 1, the effect of E_I will be greater than E_{II}. In other words, by adjusting β it can be decided whether the focus will be on fast recovery of the damaged lines or on minimizing the cost of repair crew dispatched to damaged locations. The F_β-score can be written as a percentage value, where F_β-score = 100% means the classifier has no error, and F_β-score = 0% means the classifier cannot correctly identify any of the line outages.

3.2.4. Incompleteness in Data

The last gasp signals may not be sent out or received to be utilized in the proposed method. In this respect, the performance of the proposed method under the presence of incomplete signals will be investigated by preprocessing them in a step named data preprocessing. The incomplete signals could be categorized into two different classes: semi-incomplete and pure-incomplete. In what follows, these two classes of incomplete signals will be first defined and further comprehensively preprocessed to be utilized in the proposed method.

Semi-incomplete signals are such kinds of signals which their statuses could be exactly determined through the last gasp signals received by their prior and subsequent smart meters. As shown in Fig. 3.7, the last gasp signals of three smart meters connected to buses 2, 3, and 13, are not sent out or received. These incomplete signals are classified as semi-incomplete as their statuses can be exactly determined by investigating their adjacent smart meter statuses. The preprocessing step is employed to determine the semi-incomplete signals statuses. In this step, by investigating the statuses of prior and subsequent smart meters, the statuses of the missed smart meters will be specified. In Fig. 3.7, for buses 2 and 3, as the subsequent smart meters from bus 4 to bus 7 are in service, the smart meters in buses 2 and 3 are determined to be in service. For bus 13, as the prior smart meter's status in bus 12 is on outage, the smart meter in bus 13 is governed to be on outage. Thus, this class of incomplete signals, i.e., semi-incomplete, can be precisely resolved based on this preprocessing step, and further utilized in the proposed method.

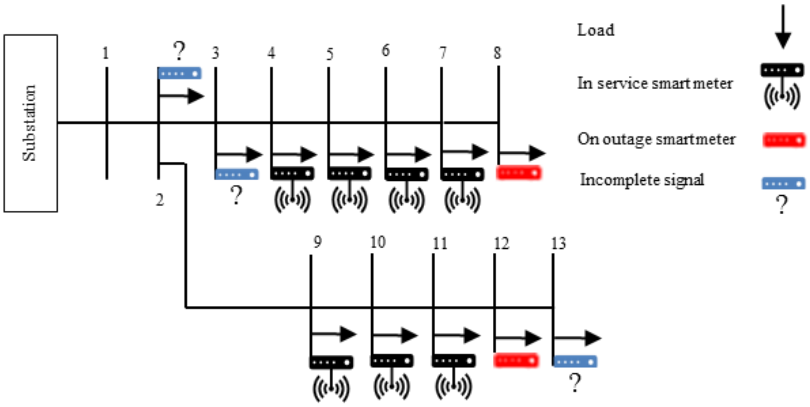


Fig. 3.7: An illustrative example of preprocessing step to determine semi-incomplete data.

On the other hand, pure-incomplete signals are that kind of smart meter statuses that are not possible to be identified via the last gasp signals received by their adjacent smart meters. As demonstrated in Fig. 3.8, the last gasp signals of two smart meters connected to buses 7 and 13 are not sent out or received, and these incomplete signals are categorized as pure-incomplete since their statuses cannot be determined by their adjacent smart meters. Regrading bus 7, as the prior smart meters are in service and the subsequent smart meter in bus 8 is on outage, the status of the smart meter is not clear. This is the case for the smart meter in bus 13, as the prior smart meters are in service and there is no evidence to define the status of the smart meter in bus 13.

The smart meter statuses associated with these pure-incomplete signals could be considered either as in service or on outage. As mentioned in section III of the dissertation, two types of error are defined. Error type I or False Negative means an on outage smart meter is not identified correctly and it leads to longer outage time in that location. Error type II or False Positive means an in-service smart meter is classified as an on outage one wrongly, which leads to sending unnecessary repair personnel to the location. In the case of pure-incomplete signals, by considering them as in service, the Error type I may be increased which means some of the damaged lines are wrongly classified as in service. On the other hand, if they are regarded to be on outage, the Error type II E_{II} may be increased, denoting some of in-service lines are wrongly classified as on outage. Nevertheless, as the proposed method prefers to restore outages in the shortest time at the expense of the extra cost of sending additional repair crew to damaged sites, the pure-incomplete signals associated with smart meters are set to be on outage. By doing this, the proposed method can achieve its quick restoration goal in the shortest possible time.

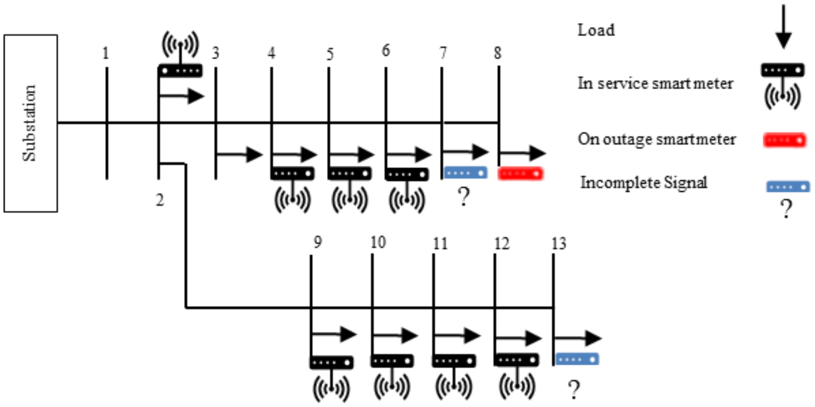


Fig. 3.8: An illustrative example of preprocess step to determine pure-incomplete data.

3.3 Numerical Studies

The proposed outage identification method is applied to both small and relatively large networks to check its accuracy and scalability. In Case 1, the proposed method is tested on the IEEE 33-bus test system. The IEEE 123-bus distribution test system is further used in Case 2 to examine the ability of the proposed method in case of a relatively large network. Calculations were done in MATLAB. Case 3 uses IEEE 123-bus test system to evaluate the effectiveness of the proposed method in case of incompleteness in the data, i.e., lost last gasp signals of smart meters.

3.3.1. IEEE 33-Bus Distribution Test System

The IEEE 33-bus radial distribution network is used to evaluate the proposed ML-SVM classifier for line outage identification. Fig. 3.9 shows the IEEE 33-bus distribution test system. The training input vector $\boldsymbol{x}_i$ is a 32-dimensional vector, corresponding to the number of reporting smart meters. The connecting lines between the buses are considered as classes in the ML-SVM classifier, so the number of the classes based on the network topology is 32.

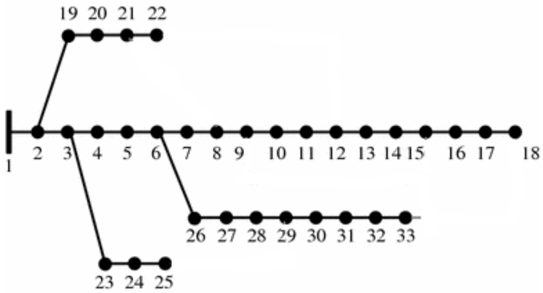


Fig. 3.9: IEEE 33-bus distribution test system.

A set of 385 scenarios based on the network topology is generated to synthesize the data, of which 80% and 20% of the generated data are respectively employed to train and test the ML-SVM classifier. For the test samples, 1517 outages are reported by the smart meters, which are correspondingly caused by 239 damaged lines. It should be noted that the mentioned numbers for outages and damaged lines are extracted from all generated scenarios associated with the test samples. The proposed multi-label scheme with linear, polynomial, and RBF kernels is performed to locate the damaged lines. Based on trial and error, the associated parameters for the polynomial kernel, i.e., γ_p, r and d are set as 1, 10, and 3, respectively. In a similar fashion, γ_r is found to be 0.1 in the RBF kernel. The goal here is to reduce E_I, rather than E_{II}, as utility companies mostly try to restore power in the shortest possible time while overlooking the potential costs. Without loss of generality, β will be considered as 2 in this study.

The mentioned kernels represent various results in terms of confusion matrix, F_2-score, and computation time as tabulated in Table 3.2. The ML-SVM classifier with the polynomial kernel correctly locates 238 damaged lines out of 239 actual damaged ones, and 2224 out of 2225 of lines are correctly classified as in service. The classifiers with the linear and RBF kernels, respectively, identify 226 and 202 damaged lines. As shown in Table 3.2, the E_I is smaller than the E_{II} for all three kernel functions, which means the proposed method focuses on restoration in the shortest time at the expense of the extra cost of sending additional repair crew to damage sites.

F_2-score and computation time are considered as two decisive factors in selecting the best kernels. The calculated F_2-scores for the ML-SVM with linear, polynomial, and

RBF kernels are calculated as 96.17%, 99.58%, and 89.62%, respectively. The polynomial kernel has the highest F_2-score and the least computation time compared to the other two kernels, so it can be considered as the most suitable classification method for the proposed problem.

Table 3.2: Comparison of various kernels of ML-SVM for 33-bus network

Kernel	**Confusion Matrix**		**F_2-Score (%)**	**Computation time (s)**
Linear	2217	13	96.17	25.7
	8	226		
Polynomial	2224	1	99.58	18.6
	1	238		
RBF	2205	37	89.62	21.7
	20	202		

Fig. 3.10 provides the graphical result of the ML-SVM classifier with a polynomial kernel for one test data as a sample. The proposed method correctly identifies lines 3-4, 21-22, and 23-24 as damaged ones. The result advocates the fact that even though numerous outages are reported by the related smart meters, only three damaged lines are the sources of these outages.

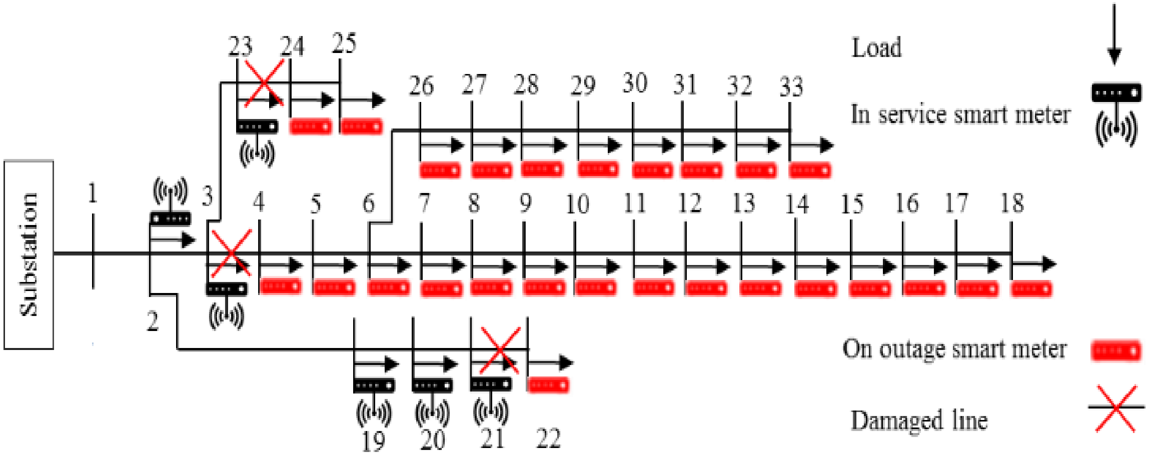


Fig. 3.10: The graphical result of ML-SVM classifier with polynomial kernel for one test data as a sample in the IEEE 33-bus.

3.3.2. IEEE 123-Bus Distribution Test System

To evaluate the scalability of the proposed method, the ML-SVM is tested on the IEEE 123-bus distribution network. As shown in Table 3.3, while the size of the network is increased, the efficiency of the proposed method is still high, and the solution is obtained in around 1 min. It should be mentioned that if the low voltage networks are included, the number of measurements would considerably increase, however, as demonstrated, the proposed method is scalable and capable of finding the solution for much larger systems in a relatively short amount of time.

Table 3.3: Comparison of various kernels of ML-SVM for 123-bus network

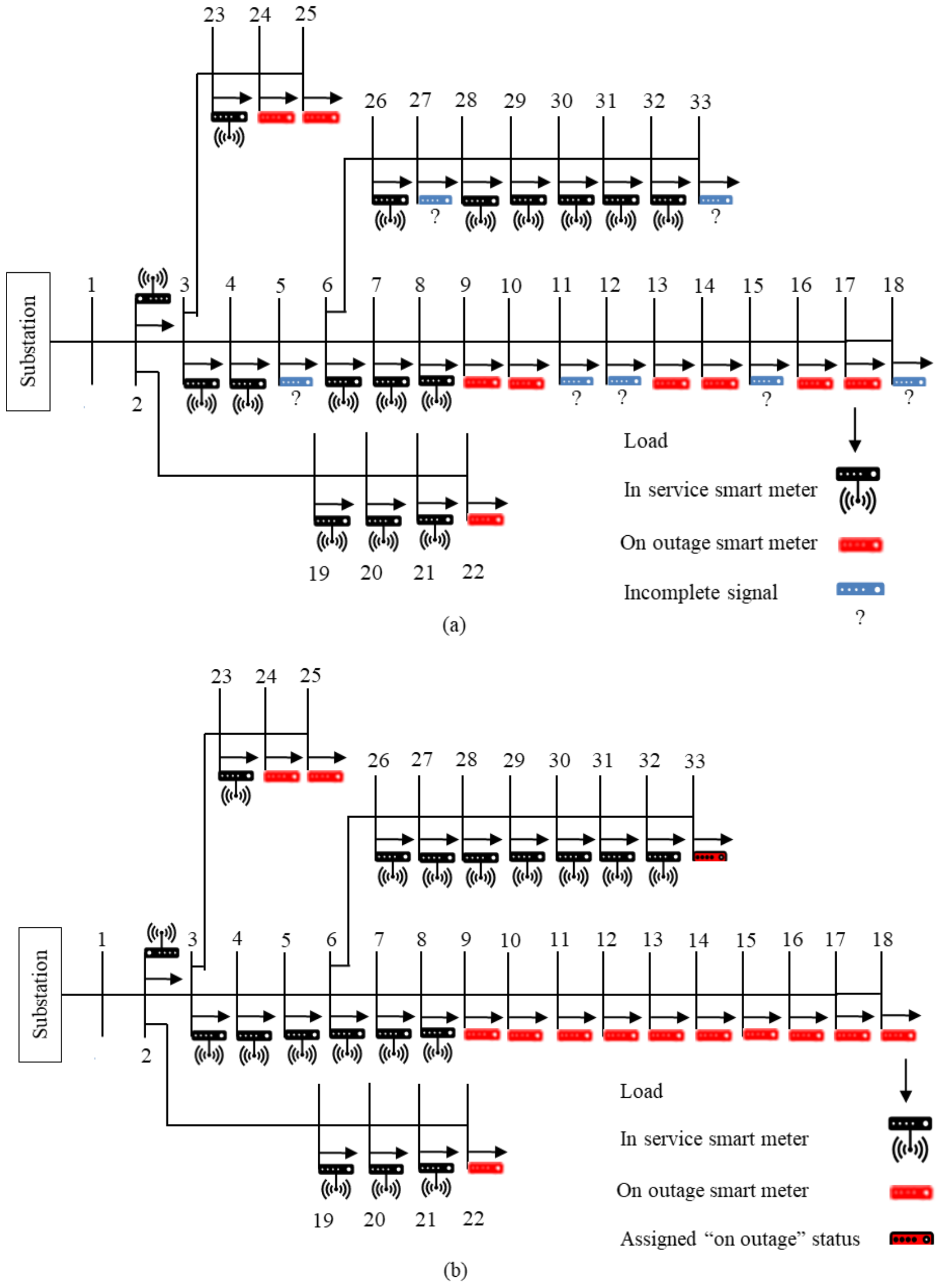
Kernel	Confusion Matrix		F_2-Score (%)	Computation time (s)
Linear	16278	125	94.91	68.3
	76	1601		
Polynomial	16322	39	98.06	58.1
	32	1687		
RBF	16211	197	90.86	60.8
	143	1529		

3.3.3. Lost Last gasp Signals

The last gasp signals may not be sent out or received to be utilized in the proposed method. In this respect, the performance of the proposed method under the presence of incomplete signals is investigated by data preprocessing on the IEEE 123-bus test system.

In this case, it is assumed that the last gasp signals of a set of random buses in the network are not received by the utility. Randomness is modeled by a uniformly distributed pseudorandom integer producer. Sensitivity analysis with respect to the percentage of lost last gasp signals in the test dataset is performed for the IEEE 123-bus test system. The percentage of lost last gasp signals changes from 10% to 50% by steps of 10%, while three kernels, namely linear, polynomial, and RBF are used for the ML-SVM classifier.

Fig. 3.11 shows an example scenario of the IEEE 33-bus test system with 20% of buses experiencing missing last gasp signals before and after the proposed preprocessing step. It should be noted that 20% of the buses in this system are roughly equal to 7 out of 33 buses.



(a)

(b)

Fig. 3.11: An illustrative example of the IEEE 33-bus test system with 20% lost last gasp signals, (a) before and (b) after the preprocessing step.

As this figure illustrates, this example contains both pure-incomplete as well as semi-incomplete data, where bus 33 contains pure-incompleteness while buses 5, 11, 12, 15, 18, and 27 are examples of semi-incompleteness. Based on the proposed methodology in this dissertation, for the example shown in the figure, six buses containing semi-incomplete data would be easily identified and replaced with their actual expected last gasp signals. In other words, as the smart meters prior to buses 11, 12, 15, and 18 are reporting outage, the corrected last gasp signals for the mentioned buses would be on outage. On the other hand, the status of buses 5 and 27 would be replaced by in-service status, as their subsequent buses are in-service ones. The only bus unidentified would be bus 33, which would be dealt by as pure-incomplete data in the proposed method. Therefore, the proposed preprocessing step reduces the incompleteness from 20% to only 3% in this example. However, as the goal of the proposed method is to identify all the potential outages in order to have a fast power recovery, "on outage" status would be assigned to the pure-incomplete data.

The results of sensitivity analysis for the IEEE 123-bus test system are demonstrated in Fig. 3.12. The percentage of lost last gasp signals changes from 10% to 50% by steps of 10%, while three kernels, namely linear, polynomial, and RBF are used for the ML-SVM classifier. As shown, in the case of lost last gasp signals, when the percentage of random buses with missing last gasp signal increase from 10% to 50%, accuracy is slightly decreased. Among the three studied kernels, RBF has the least robustness in case of missing data, while the polynomial kernel shows a high performance in handling incompleteness in data. However, this case validates that the proposed method is capable of identifying residential customers' phases with acceptable accuracy.

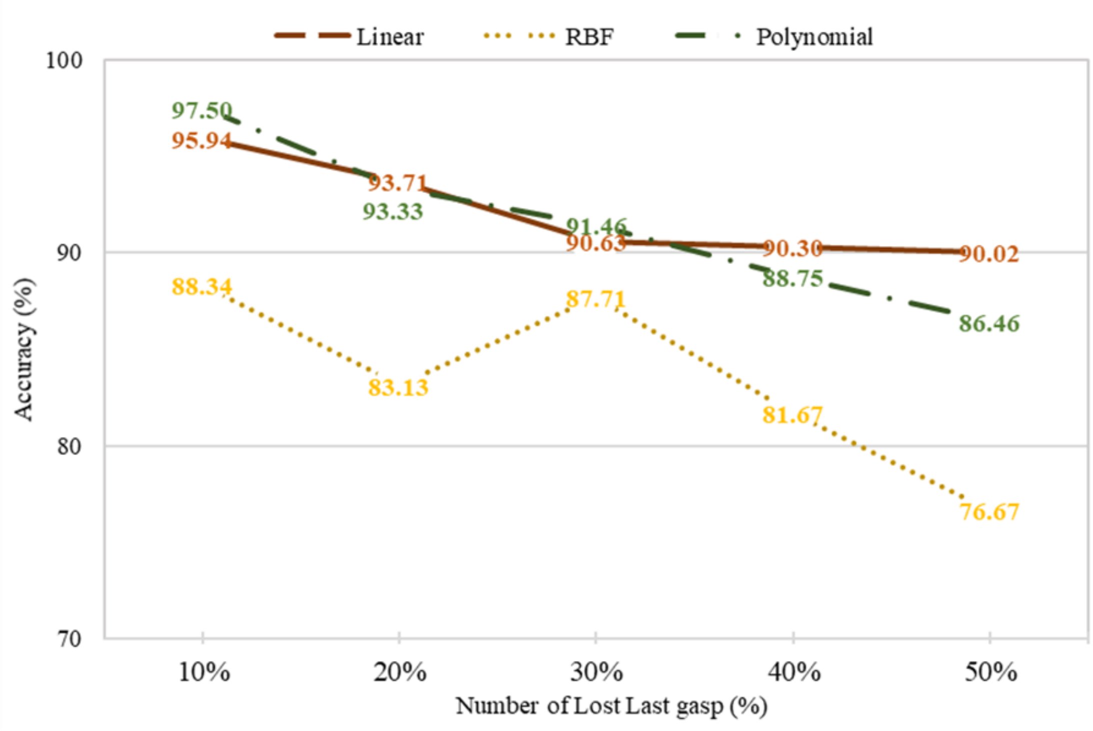


Fig. 3.12: The accuracy of the outage identification method in case of lost last gasp signals based on different kernels.

3.4 Conclusion

A novel scheme was proposed in this dissertation to identify distribution lines outages in response to extreme weather events by leveraging AMI data. The proposed method benefits from an ML-SVM classification method and by using the last gasp signals collected through AMIs, the proposed method identifies the outage locations in the distribution network. The proposed ML-SVM solution consists of four steps; data synthesis, training, testing, and evaluation. Through data synthesis, the historical network topology and the outage status of smart meters are used to label the actual damaged lines, while the second step utilizes a large portion of the synthesized data and several binary SVMs to train a multi-label classifier and generate the associated SVM parameters for the test step. The rest of the synthesized data, along with the generated SVM parameters, are employed in the evaluation step.

The effective and acceptable performance of the proposed scheme was validated through numerical simulations by considering small and large distribution networks. In addition, the proposed method was evaluated in the case of lost last gasp signals. Numerical simulations demonstrated that the proposed AMI-based distribution outage identification was fast, accurate, and could efficiently identify the line outages. Utility companies can reap the benefits of this intelligent method to accelerate the process of grid response and recovery and consequently, decrease the associated outage durations and costs.

References

[1] M.H. Wen, R. Arghandeh, A. von Meier, K. Poolla, and V. O. Li, “Phase identification in distribution networks with micro-synchrophasors,” IEEE Power & Energy Society General Meeting, Denver, CO, July 2015.

[2] A. and Shuster, M., “Electric company smart meter deployments: foundation for a smart grid (2021 update),” The Institute for Electric Innovation (IEI): Washington, DC, USA, 2021.

[3] H. Pezeshki and P.J. Wolfs, “Correlation based method for phase identification in a three phase LV distribution network,” 22nd Australasian Universities Power Engineering Conference (AUPEC), Bali, Indonesia, Sept. 2012.

[4] J. Zhu, M.Y. Chow, and F. Zhang, “Phase balancing using mixed-integer programming [distribution feeders],” IEEE transactions on power systems, vol. 13, no. 4, pp. 1487-1492, 1998.

[5] D. K. Chembe, “Reduction of power losses using phase load balancing method in power networks,” In Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp. 20-22, 2009.

[6] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst, “Phase identification of smart meters by clustering voltage measurements,” Power Systems Computation Conference, Dublin, Ireland, June 2018.

[7] Z. S. Hosseini, M. Mahoor, and A. Khodaei, “AMI-Enabled Distribution Network Line Outage Identification via Multi-Label SVM,” IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 5470-5472, Sept. 2018.

[8] C. Lueken, P. M. Carvalho, and J. Apt, “Distribution grid reconfiguration reduces power losses and helps integrate renewables,” Energy Policy, vol. 48, pp. 260-273, Apr. 2012.

[9] F. Melo, C. Cândido, C. Fortunato, N. Silva, F. Campos, and P. Reis, “Distribution automation on LV and MV using distributed intelligence,” 22nd International Conference and Exhibition on Electricity Distribution (CIRED), Stockholm, Sweden, June 2013.

[10] C. S. Chen, T. T. Ku, and C. H. Lin, “Design of phase identification system to support three-phase loading balance of distribution feeders,” IEEE Industrial and Commercial Power Systems Technical Conference, Baltimore, MD, May 2011.

[11] L. Marrón, X. Osorio, A. Llano, A. Arzuaga, and A. Sendin, “Low voltage feeder identification for smart grids with standard narrowband PLC smart meters,” IEEE International Symposium on Power Line Communications and Its Applications (ISPLC), Johannesburg, South Africa, Mar. 2013.

[12] R. E. Brown, “Electric power distribution reliability,” New York: Marcel Dekker, 2008.

[13] C.-C. Liu, “Distribution systems: reliable but not resilient?” IEEE Power Energy Mag., vol. 13, no. 3, pp. 93–96, May 2015.

[14] Y. Jiang, C. C. Liu, M. Diedesch, E. Lee, A. K. Srivastava, “Outage management of distribution systems incorporating information from smart meters,” IEEE Transactions on Power Systems, vol. 31, no. 5, pp. 4144-4154, Sep. 2016.

[15] F. C. Trindade, W. Freitas, J. C. Vieira, “Fault location in distribution systems based on smart feeder meters,” IEEE transactions on Power Delivery, vol. 29, no.1, pp. 251-260, Feb. 2014.

[16] R. J. Campbell, “Weather-related power outages and electric system resiliency,” Washington, DC: Congressional Research Service, Library of Congress, Aug. 2012.

[17] R. F. Ghajar, R. Billinton, “Economic costs of power interruptions: a consistent model and methodology,” International Journal of Electrical Power & Energy Systems, vol. 28, no.1, pp. 29-35, Jan. 2006.

[18] K.J. Caird, General Electric Co, “Meter phase identification,” U.S. Patent 8,143,879, 2010.

[19] A.R. Kolwalkar, H.W. Tomlinson, B. Sen, J. E. Hershey, and G. P. Koste, General Electric Co, “Power meter phase identification,” U.S. Patent Application 12/782,530, 2011.

[20] J. D. Watson, J. Welch, and N. R. Watson, “Use of smart-meter data to determine distribution system topology,” The Journal of Engineering (IET), vol. 2016, no. 5, 2016.

[21] B. K. Seal and M. F. McGranaghan, “Automatic identification of service phase for electric utility customers,” IEEE Power and Energy Society General Meeting, San Diego, CA, July 2011.

[22] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, “Smart meter data analytics for distribution network connectivity verification,” IEEE Transactions on Smart Grid, vol. 6, no. 4, pp. 1964-1971, 2015.

[23] F. Olivier, D. Ernst, and R. Fonteneau, “Automatic phase identification of smart meter measurement data,” Proc. of CIRED, vol. 2017, no. 1, 2017.

[24] T. A. Short, “Advanced Metering for Phase Identification, Transformer Identification, and Secondary Modeling,” IEEE Transactions on Smart Grid, vol. 4, no. 2, pp. 651-658, 2013.

[25] W. Wang, N. Yu, and Z. Lu, “Advanced Metering Infrastructure Data Driven Phase Identification in Smart Grid,” The Second International Conference on Green Communications, Computing and Technologies (GREEN 2017), Rome, Italy, Sept. 2017.

[26] L. Blakely, M. J. Reno, and W.C. Feng, “Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries,” IEEE Power and Energy Conference at Illinois, Champaign, IL, Feb. 2019.

[27] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, “Phase identification in electric power distribution systems by clustering of smart meter data,” IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, Dec. 2016.

[28] M. Dilek, R. P. Broadwater, and R. Sequin, "Phase prediction in distribution systems," IEEE Power Engineering Society Winter Meeting, Conference Proceedings, New York, NY, Jan. 2002.

[29] V. Arya, D. Seetharam, S. Kalyanaraman, K. Dontas, C. Pavlovski, S. Hoy, and J. R. Kalagnanam, "Phase identification in smart grids" IEEE International Conference on Smart Grid Communications (SmartGridComm), Brussels, Belgium, Oct. 2011.

[30] V. Arya, T.S. Jayram, S. Pal, and S. Kalyanaraman, "Inferring connectivity model from meter measurements in distribution networks," The fourth international conference on Future energy systems, Berkeley, CA, May 2013.

[31] A. Brint, G. Poursharif, M. Black, and M. Marshall, "Using grouped smart meter data in phase identification," Computers & Operations Research, vol. 96, pp. 213-222, 2018.

[32] M. Xu, R. Li, and F. Li, "Phase identification with incomplete data," IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 2777-2785, 2018.

[33] J. P. Satya, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "Identifying Topology of Low Voltage Distribution Networks Based on Smart Meter Data," IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 5113-5122, 2017.

[34] Z. S. Hosseini, A. Khodaei, and A. Paaso, "Machine Learning-Enabled Distribution Network Phase Identification," IEEE Transactions on Power Systems, vol. 36, no. 2, pp. 842-850, Mar. 2021.

[35] U.S. Department of Energy, "Advanced Metering Infrastructure and Customer Systems: Results from the Smart Grid Investment Grant Program," Office of Electricity Delivery and Energy Reliability, Sep. 2016."

[36] M. Dash and H. Liu, "Feature selection for clustering," Pacific-Asia Conference on knowledge discovery and data mining, Springer, Berlin, Heidelberg, April 2000.

[37] R. Elankavi, R. Kalaiprasath, and D. R. Udayakumar, "A fast clustering algorithm for high-dimensional data," International Journal of Civil Engineering and Technology (IJCIET), vol. 8, no. 5, pp.1220-1227, 2017.

[38] M. R. Patel, “Wind and solar power systems: design, analysis, and operation,” CRC press, 2005.

[39] R. Chedid, S. Rahman, "Unit sizing and control of hybrid wind-solar power systems," IEEE Transactions on Energy Conversion, vol. 12, pp. 79-85, 1997.

[40] A. Majzoobi, A. Khodaei, and S. Bahramirad, “Capturing distribution grid-integrated solar variability and uncertainty using microgrids,” IEEE PES General Meeting, Chicago, IL, 2017.

[41] M. Mahoor, N. Iravani, S. M. Salamati, A.Aghabali, AshkanRahimi- Kian, “Smart Energy Management for a Micro-grid with Consideration of Demand Response Plans, ” IEEE Smart Grid Conference, Tehran, Iran, Nov. 2013.

[42] Z. S. Hosseini, A. Khodaei, S. Bahramirad, L. Zhang, A. Paaso, M. Lelic, and D. Flinn, “Levelized Cost of Energy Calculations for Microgrid-Integrated Solar-Storage Technology,” IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Chicago, IL, Apr. 2020.

[43] P. M. Corrigan and G. T. Heydt, "Optimized dispatch of a residential solar energy system," IEEE North American Power Symposium (NAPS), Las Cruces, NM, Sep. 2007.

[44] A. J. Black, "Financial payback on California residential solar electric systems," Solar Energy, vol. 77, no. 4, pp. 381-388, Oct. 2004.

[45] SEIA, “Solar Market Insight 2015 Q4,” Solar Energy Industries Association research report, [Online]. Available: http://www.seia.org/research-resources/solar-market-insight-2015-q4.

[46] Z. S. Hosseini, M. Mahoor, and A. Khodaei, “Battery Swapping Station as an Energy Storage for Capturing Distribution-Integrated Solar Variability,” IEEE North American Power Symposium (NAPS), Fargo, ND, Sept. 2018.

[47] H. Sadeghian, M. H. Athari, and Z. Wang, "Optimized Solar Photovoltaic Generation in a Real Local Distribution Network," IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, Apr. 2017.

[48] S. H. Elyas, H. Sadeghian, H. O. Alwan, and Z. Wang, "Optimized Household Demand Management with Local Solar PV Generation," IEEE North American Power Symposium (NAPS), Morgantown, WV, Sept. 2017.

[49] H. Sadeghian and Z. Wang, " Decentralized Demand Side Management with Rooftop PV in Residential Distribution Network," IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, Apr. 2017.

[50] Z. S. Hosseini, A. Khodaei, E.A. Paaso, M.S. Hossan, and D. Lelic, "Dynamic Solar Hosting Capacity Calculations in Microgrids", CIGRE US National Committee (USNC) Grid of the Future (GOTF) Symposium, Reston, VA, Oct. 2018.

[51] SEIA, “Solar Market Insight 2019 Q3,” Solar Energy Industries Association research report, [Online]. Available: https://www.seia.org/research-resources/solar-market-insight-report-2019-q3.

[52] Commission for Energy Regulation (CER), “CER Smart Metering Project - Electricity Customer Behavior Trial, 2009-2010 [dataset],” 1st Edition, Irish Social Science Data Archive, 2012. SN: 0012-00. www.ucd.ie/issda/CER-electricity.

[53] MATLAB, “MATLAB. 9.7.0.1190202 (R2019b),” Natick, Massachusetts: The MathWorks Inc., 2018.

[54] GAMS Development Corporation, “General Algebraic Modeling System (GAMS) Release 26.1.0,” Fairfax, VA, USA, 2019.

[55] NREL, “NREL's Solar Power Data for Integration Studies [dataset]”, Accessed on Aug. 2021, www.nrel.gov/grid/solar-power-data.html.

[56] The Economist, “Power outages like the one in Texas are becoming more common in America,” Available online: https://www.economist.com/graphic-detail/2021/03/01/power-outages-like-the-one-in-texas-are-becoming-more-common-in-america.

[57] Y. Zhao, J. Chen, and H. V. Poor, “Learning to infer: a new variational inference approach for power grid topology identification,” IEEE Statistical Signal Processing Workshop (SSP), Palma de Mallorca, Spain, Jun. 2016.

[58] Y. Zhao, J. Chen, and H. V. Poor, “Efficient neural network architecture for topology identification in smart grid,” IEEE Global Conference on Signal and Information Processing (GlobalSIP), Washington, DC, Dec. 2016.

[59] G. Tsoumakas, I. Katakis, “Multi-label classification: An overview,” International Journal of Data Warehousing and Mining 3.3, 2006.

[60] B. Schölkopf, C. J. Burges, A. J. Smola, editors, “Advances in kernel methods: support vector learning,” MIT press, 1999.

[61] H. Y. Huang and C. J. Lin, “Linear and kernel classification: When to use which?” 2016 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, pp. 216-224, June 2016.

[62] S. Maji, A. C. Berg, and J. Malik, “Classification using intersection kernel support vector machines is efficient,” IEEE conference on computer vision and pattern recognition, pp. 1-8, June 2008.

[63] D. X. Zhou and K. Jetter, “Approximation with polynomial kernels and SVM classifiers,” Advances in Computational Mathematics, 25(1), pp.323-344, 2006.

[64] J. Rousu, C. Saunders, S. Szedmak, J. Shawe-Taylor, K. P. Bennett, and E. Parrado-Hernández, “Kernel-based learning of hierarchical multilabel classification models,” Journal of Machine Learning Research, 7(7), 2006.

[65] Q. Chang, Q. Chen, and X. Wang, “Scaling Gaussian RBF kernel width to improve SVM classification,”IEEE International Conference on Neural Networks and Brain, vol. 1, pp. 19-22, Oct. 2005.

Appendix: List of Publications and Awards

- **Journal Papers:**

1. **Zohreh S. Hosseini**, A. Khodaei, and A. Paaso, "Machine Learning-Enabled Distribution Network Phase Identification," IEEE Transactions on Power Systems, 36(2), pp.842-850, Mar. 2021.
2. M. Mahoor, **Zohreh S. Hosseini**, A. Khodaei, A. Paaso, and D. Kushner, "State-Of-The-Art in Smart City Streetlight Systems: A Review," IET Smart Cities, 2(1), pp.24-33, Apr. 2020.
3. M. Mahoor, **Zohreh S. Hosseini**, and A. Khodaei, "Least-Cost Operation of a Battery Swapping Station with Random Customer Requests," Energy, vol. 172, pp.913-921, Apr. 2019.
4. **Zohreh S. Hosseini**, M. Mahoor, and A. Khodaei, "AMI-Enabled Distribution Network Line Outage Identification," IEEE Transactions on Smart Grid, 9(5), pp.5470-5472, Sept. 2018.

- **Conference Papers:**

1. **Zohreh S. Hosseini**, A. Khodaei, S. Bahramirad, L. Zhang, A. Paaso, M. Lelic, and D. Flinn, "Levelized Cost of Energy Calculations for Microgrid-Integrated Solar-Storage Technology," IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Chicago, IL, Apr. 2020.
2. **Zohreh S. Hosseini**, A. Khodaei, A. Paaso, M. S. Hossan, and M. Lelic, "Dynamic Solar Hosting Capacity Calculations in Microgrids," CIGRE US National Committee (USNC) Grid of the Future (GOTF) Symposium, Reston, VA, Oct. 2018.
3. B. Nguyen, **Zohreh S. Hosseini**, and D. W. Gao, "Analysis of Pricing Trends and Grid Parity of Photovoltaic Systems," CIGRE US National Committee (USNC) Grid of the Future (GOTF) Symposium, Reston, VA, Oct. 2018.

4. **Zohreh S. Hosseini**, M. Mahoor, and A. Khodaei, "Battery Swapping Station as an Energy Storage for Capturing Distribution-Integrated Solar Variability," IEEE North American Power Symposium, Fargo, ND, Sept. 2018 **[Best paper award]**.
5. M. Mahoor, **Zohreh S. Hosseini**, A. Khodaei, and D. Kushner, "Electric Vehicle Battery Swapping Station," CIGRE US National Committee (USNC) Grid of the Future (GOTF) Symposium, Cleveland, OH, Oct. 2017.
6. M. Mahoor, A. Majzoobi, **Zohreh S. Hosseini**, and A. Khodaei, "Leveraging Sensory Data in Estimating Transformer Lifetime," IEEE North American Power Symposium (NAPS), Morgantown, WV, Sept. 2017.

- **Awards:**

1. "DU Graduate Education Dissertation Fellowship" for Spring 2021.
2. "DU Graduate Education Doctoral Fellowship for Inclusive Engagement" for 2020-2022.
3. "DU Graduate Studies Doctoral Fellowship" for 2018-2019.
4. 2020 Colorado Winner of the National Center for Women & Information Technology (NCWIT) Award for Aspirations in Computing for the project named "A Modified Machine Learning-based Clustering Approach for Residential Customers' Phase Identification in Power Systems".
5. Finalist of the national 2020 National Center for Women & Information Technology (NCWIT) Collegiate Award for Aspirations in Computing for the project named "A Modified Machine Learning-based Clustering Approach for Residential Customers' Phase Identification in Power Systems".
6. Best Paper Award for the paper titled "Battery Swapping Station as an Energy Storage for Capturing Distribution-Integrated Solar Variability" at the North American Power Symposium in 2018.