Doctoral Dissertation
Doctoral Program in Energy Engineering (32nd Cycle)

Data Challenges and Data Analytics Solutions for Power Systems

Yang Zhang
* * * * * *

Supervisors
Prof. Ettore Bompard

Doctoral Examination Committee:
Prof. A.B. , Referee, University of….
Prof. C.D. , Referee, University of…
Prof. E.F. , Referee, University of….
Prof. G.H. , Referee, University of…
Prof. I.J. , Referee, University of….

Politecnico di Torino
April 15, 2020

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

…….....

Yang Zhang

Turin, April 15, 2020

Summary

The present work of thesis mainly presents some innovative data-driven applications to deal with the emerging data challenges in power system.

With the overwhelming trend of digitalization in power system, critical challenges and opportunities are emerging in the modern industrial systems. In the first chapter, the big data analytics are introduced with corresponding applications in smart grids. With huge amount of data from electricity network, meteorological information system, geographical information system etc., many benefits can be brought to the existing network and import the customer service as well as the social welfare in the era of big data.

In distribution network, there are a large number of feeders spread in a wide area with a diversity of structural features. Chapter 2 discusses the possibility to group thousands of feeders with a data-driven method and try to identify the representative feeders for a better evaluation of the performance with regard to the number of outages. Two clustering algorithms are applied on the mixed data with the similar analysis results.

In Chapter 3, the data-driven solutions to the problems relating to the outage predictions in medium-voltage distribution network is presented. Annual and monthly number of outages are predicted based on improved grey theory. Further more, the daily outages prediction is discussed as a binary classification problem with support vector machine.

Facing the challenges of limited channel bandwidth in advanced metering infrastructures, the consumption records with partial missing information are utilized in Chapter 4 to figure out the abnormal users in power system. By extracting the importance index through regression models, anomalies from the energy consumption patterns are successfully uncovered.

Finally, the Monte Carlo simulation is introduced to evaluate the high-impact low-probability events in the local distribution network in Chapter 5. Since there are considerable increase of outages in the summer, heat waves are regarded as a severe threat to the power system security. A detailed analysis between the temperature and number of outages is firstly discussed as the basis of probabilistic simulation. The savings of expenditures due to a decrease of repetitive faults after investment could be estimated via calculation from thousands of scenarios.

Acknowledgment

First of all, I would like to thank my supervisor, professor Ettore Bompard. During the last few years, he accompanied and encouraged me to get out of the comfort zone and face new challenges to do a promising job. Thank you very much and I would forever be grateful for all his support.

A special thank is for my parents, who supported me during my student career with unconditional love.

A huge thank is for professor Andrea Mazza, who gave me a lot of guidance on the road of scientific research and always inspired me with innovative ideas.

Furthermore, I would like to thank all my colleagues and friends: Shaghayegh and Francesco, who are enrolled in PoliTo at the same year with me and accompanied me the longest time; Renjian, my senior and also friend; Abouzar, who is always kind and optimistic; Giulia, Carmelo, Mahmood and all the other colleagues at the same department.

Finally, I would like to thank all the people I met in these few years and helped me to deal with the problems both in electrical engineering world and real life.

I would like to dedicate this thesis to my loving parents

Contents

List of Tables

List of Figures

Chapter 1

Data analytics in power system: an overview

1.1 Big data in smart grid

With the fast development of digital technology and cloud computing, more and more data are produced through digital equipment and sensors, such as smart phones, computers, advanced measuring infrastructures, etc., as well as through human activities and communications. For instance, the size of data on the internet is now measured in exabytes (1018) and zettabytes (1021) [1]. Rational, effective and efficient analysis of these data brings huge value and benefit to our daily life and company activities. However, the collected data are mounting at an exponential growth, and the structure of them is also becoming much more complicated. The processing and analysis method of these large volume data is a new challenge but opportunity at the beginning of this century with the concept of "big data" [2][3].

Although big data is a newly-appeared term, the concept of discovering valuable information from massive collected data in commercial operation as aiding knowledge for business decision has already been proposed in 1989 by Howard Dresner as "business intelligence" (BI) [4]. The trend of internet revolution and ubiquitous information acquisition devices successfully reduce the cost of data collection, while the huge amount and complex structure challenge the capability of traditional data analytics techniques.

In power grid, the traditional fossil fuels are facing the problem of depletion and the de-carbonization demands the power system to reduce the carbon emission. Smart grid and super grid are effective solutions to accelerate the pace for electrification of human society with high penetration of renewable energy sources [5]. Although the rising awareness of sustainable development have become the impetus to the utilization of renewable energy sources, the intermittent characteristics of wind and photovoltaic energies bring huge

challenges to the safe and stable operation in a low inertia power system[6], [7]. The data analytics based renewable energy forecasting methods are a hot research topic for a better regulation and dispatch planning in such cases. Traditional electricity meters in distribution systems only produce a small amount of data which can be manually collected and analyzed for billing purpose. While the huge volume of data collected from two-way communication smart grids at different time resolutions in nowadays need advanced data analytics to extract valuable information not only for billing information but also the status of the electricity network. For example, the high-resolution user consumption data can also be used for customer behavior analysis, demand forecasting and energy generation optimization. Predictive maintenance and fault detection based on the data analytics with advanced metering infrastructure are more crucial to the security of power system [8].

Thus, the great progress of information and communication technology (ICT) provides a new vision for engineers to perceive and control the traditional electrical system and makes it smart. An embedded information layer into the energy network produces huge volume of data, including measurements and control instructions in the grid for collection, transmission, storage and analysis in a fast and comprehensive way. It also brings a lot of opportunities and challenges to the data analysis platform. This paper is to discuss the concepts of data analysis and their applications in smart grids. The intent of this paper is three-fold. First the potential data collected with advanced metering infrastructure in smart grid are discussed. Next, the paper briefly reviews the concepts of data analytics and the popular techniques. Finally, the paper illustrates the detailed applications of data analytics in smart grid.

1.1.1 Concept of big data

The definition of big data is not very clear and uniform at present. But there is a consensus among different descriptions: this is an emerging technical problem brought by a dataset of large volume, various categories and complicated structures which needs novel framework and techniques to excavate useful information effectively. Therefore, the definition of big data depends on the ability of data mining algorithms and the corresponding hardware equipment to deal with large volume datasets [9]. It is a relative concept instead of an absolute definition. The big data can be understood as amount of data beyond technology's capability to store, manage and process efficiently in [10] as the data size increasing along with the evolvement of ICT technologies.

1.1.2 Concept of smart grids

Smart grid is the power system embedded with an information layer that allows for two-way communication between the central controllers and local actuators as well as logistic units to respond digitally to urgent situations of physical elements or quickly changing of electric demand. The E.U. defined the

smart grid as “electricity networks that can intelligently integrate the actions of all users connected to it – generators, consumers and those that do both – in order to efficiently deliver sustainable, economic and secure electricity supplies [11]. The U.S. defined the smart grid of future in a similar way that incorporates the digital technology to improve reliability, security and efficiency of the electric system through information exchange, distributed generation and storage resources for a fully automated power delivery network [12].

Compared with traditional power systems, the widespread application of distributed generators under the call of green energy resources is shaking the hegemony position of large-scale centralized power plants, which makes the conventional centralized control strategy less effective due to the unidirectional power flow. Connection of small-scale power generations (typically in the range of 3kW to 10kW) to the public distribution grid requires two-directional operation and control of distribution grids. Faced with the challenges of more complicated control and protection strategies, the conventional electro-mechanical electric grid is supposed to be enhanced with the help of innovations in the digital information and telecommunications network to overcome the cost from power outages and power quality disturbances as billions of dollars annually [13].

Normally, the smart grid can be assessed with a Smart Grid Architecture Model (SGAM), which is a 3-dimensional framework that merges domains, zones and layers together. The conventional structure of power system can be found in the domains as generation, transmission, distribution, DER (Distributed Energy Resources) and customer premises. The zones which present the layout of power system management are composed of market, enterprise, operation, station, field and process. On top of the first two dimensions, the layout of interoperability layers includes the component, communication, information, function and business layers. SGAM as an architectural overview can be used to find the limitations and commonalities of existing smart grid standards [14].

1.1.3 Big data characteristics in smart grid

The characteristics of big data in smart grid is also in accordance with the universal 5V big data model in many researches [15] as below:

Volume – refers to the vast amount of data generated, which makes data sets too large to store and analyse using traditional database technology. The possible solution to this problem is the distributed systems to store data in different locations, connect them by networks and bring them together by software. In smart grid the widespread application of smart meter and advanced sensor technology provide huge amount of data.

Velocity – refers to the speed at which new data is generated and the speed at which data moves around. The requirements for real-time exchange of data is increasing. With a sampling rate of 4 times per hour, 1 million smart meters installed in the smart grid would result in 35.04 billion records, equivalent to 2920 Tb data in quantification [16]. The following Table 1-1 indicates the amount of

records from smart meters in a year under various collection frequency with the assumption of 1 million devices and a 5 KB record per collection.

Table 1-1 Quantification of collected data in different sampling rates [17]

Collection Frequency	1/day	1/hour	1/30 min	1/15 min
Records	365 million	8.75 billion	17.52 billion	35.04 billion
Volume	1.82 TB	730 TB	1460 TB	2920 TB

Variety – refers to the types of data we can now use. In the past, we focus on structured data that neatly fits into tables or rational databases such as financial or meteorological data. With big data technology, we need to handle different types of unstructured data including messages, social media conversations, digital images, sensor data, video or voice recordings, and bring them together with more traditional, structured data. According to the extensive data sources in smart grid as shown in Figure 1.1, the formats and dimensions of data are diverse in structure.

Veracity – refers to the messiness or trustworthiness of the data. The quality and accuracy are less trustworthy with such large amount of big data, which challenge the outcome data analysis. Errors of measurements in smart grid may exist due to the imperfections in devices or mistakes in data transmission. The secure and efficient power system operation relays on the data assessment and state estimation.

Value – refers to our ability to extract valuable information from the huge amount of data and derive a clear understanding of the value it brings. The larger the amount of data is, the lower the density of valuable information will be. With the improvement of intelligent devices adopted in smart grid, more and more value of big data analytics is revealed according to the various applications.

1.1.4 Data Sources in smart grid

As an intelligent system of both energy and information, smart grid is the abundant source of information, which covers the data from process of electricity generation, transmission, distribution and consumption. These data include the electrical information from distribution stations, distribution switch stations, electricity meters, and non-electrical information like marketing, meteorological as well as reginal economic data as shown in Figure 1.1 [18]. Collection and analysis of them provide essential help in scheduling of power plants, operation of subsystems, maintenance for vital power equipment and business behavior in marketing.

The data sources mentioned above can be sorted into three categories: measurement data, business data and external data [19]. Most of the operation parameters in power system are measured through installed sensors and smart meters, indicating the system's current and historical status [16] [20]. The weather conditions and social events like festivals are the external data that cannot be measured from smart meters but have an impact on the operation and planning in

power system. The business data mainly includes the marketing strategies and rivals' behaviors.

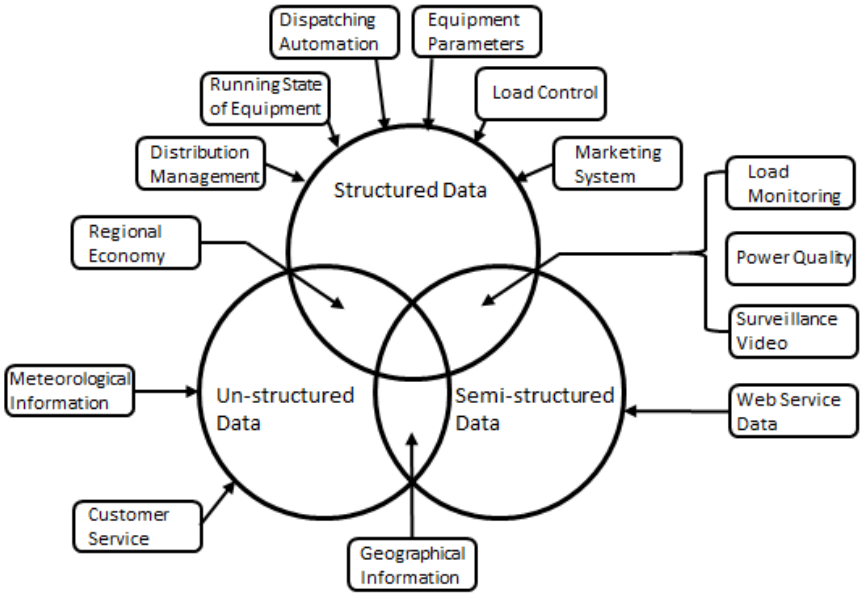


Figure 1.1 Data sources of power grid

1.1.5 Data collection techniques in smart grid

In smart grid, the data are collected and transmitted with help of smart meters which provide energy related information to both the utility company (or DSO) and customers. For the energy consumption of residential customers, the number of smart meter readings for a large utility company is expected to rise from 24 million a year to 220 million per day [16]. As an emerging component in electricity market and smart grid, electric vehicles (EVs) and plug-in hybrid EVs (PHEVs) have seen a growing popularity with the movement of electrification in transportation sector and progress of artificial intelligence. To control the normal operation status of the distribution system, DSO traditionally relies on the measurements in the primary substation, at the beginning of each MV feeder, where the protection systems are normally installed. The current magnitude information is also needed for the automatic on-load tap changer in HV/MV transformers for voltage regulation. The measurements of a typical smart meter include the node voltage, feeder current, power factor, active and reactive power, energy over a period, total harmonic distortion as well as load demand, etc. The intelligent devices for data collection in smart grid are listed as Table 1-2

Table 1-2Intelligent data collection devices in smart grid

Intelligent device	Technology	Application
Advanced metering infrastructure (AMI)	Integration of smart meters, data management systems and communication networks to provide bidirectional communication between customers and utilities.	Remote meter configuration, dynamic tariffs, power quality monitoring and local control

Phasor measurement unit (PMU)	Real-time measurements (30 to 60 samples/second) of multiple remote points with a common time source for synchronization	Electrical waves measurement of power grid
Wide area monitoring system (WAMS)	An application server to deal with the incoming information from PMUs	Dynamic stability of the grid
Remote terminal unit (RTU)	A microprocessor-controlled device that transmitting telemetry data	Information collection of system operation status
Supervisory control and data acquisition (SCADA)	Both manual and automatic	System monitoring, event processing and alarm
Intelligent electronic device (IED)	Monitoring and recording status changes in the substation and outgoing feeders	Combination of different relay protection functions with measurement, recording and monitoring

1.1.6 Data communication techniques in smart grid

The communication infrastructure of the smart grid is composed of three types of networks: home area network (HAN), neighborhood area network (NAN) and wide area network (WAN) as shown in Figure 1.2 [21]. The functions and characteristics of the above communication infrastructures are summarized in Table 1-3.

Table 1-3 Summary of communication infrastructure in smart grid

Type of network	Function	Characteristic
HAN	Enabling the communication among smart home or office devices and smart meters for local energy management	Deployed at house or small office with a relatively low transmission data rate (less than 1 Kbps)
NAN	Consisting of several HANs for energy consumption data aggregation and storage at load data center (LDC)	Deployed within area of hundreds of meters with up to 2Kbps
WAN	Enabling the communication of all smart grid's components	Deployed within tens of kilometers with high data transmission capability up to few Gbps

Basic types of communication technologies for smart meters include wired and wireless infrastructures. The wireless communication technology allows the data center to gather measurement information from smart meters with low costs and simple connections while it may face the electromagnetic problem. Power line communication (PLC) is a wired communication technology by add a modulated carrier signal to the power cables and already successfully implemented in power

system. The existing communication technology include ZigBee, WALN, cellular communication, WiMAX, PLC, etc. [21].

As one of the first countries for smart metering infrastructure development, Italy has deployed smart meter to nearly all the customers with the PLC technology to transfer smart meter data to the nearest data concentrator located in the MV/LV substation. Then these data are sent to the DSO's data centers for recording and data analysis. There are around 30 million meters and 400,000 secondary substation concentrators installed [22].

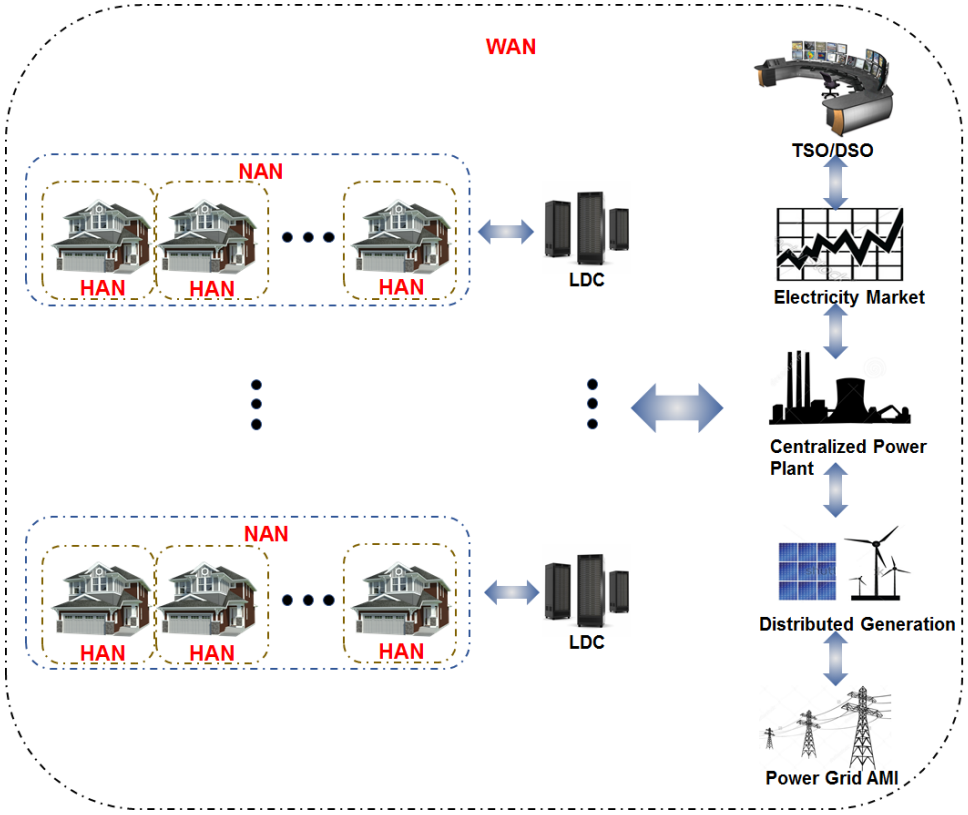


Figure 1.2 Smart grid communication infrastructure

Basic types of communication technologies for smart meters include wired and wireless infrastructures. The wireless communication technology allows the data center to gather measurement information from smart meters with low costs and simple connections while it may face the electromagnetic problem. Power line communication (PLC) is a wired communication technology by add a modulated carrier signal to the power cables and already successfully implemented in power system. The existing communication technology include ZigBee, WALN, cellular communication, WiMAX, PLC, etc. [21].

As one of the first countries for smart metering infrastructure development, Italy has deployed smart meter to nearly all the customers with the PLC technology to transfer smart meter data to the nearest data concentrator located in the MV/LV substation. Then these data are sent to the DSO's data centers for recording and data analysis. There are around 30 million meters and 400,000 secondary substation concentrators installed [22].

1.2 Data analysis techniques

The most important stage of the big data processing system is data analysis, which is the basis for discovering valuable information and supporting the decision-making [23], [24]. There are several similar concepts relevant to data analysis listed in Table 1-4.

Table 1-4 Concepts related to data analysis

Concept	description
Statistics	The study of data collection, analysis and interpretation with mathematics methods which may discover potential relations based on some hypothesis
Machine learning	A kind of technique for understanding the law in the data as well as extracting useful information with the help of computers automatically instead of humanity
Data mining	Computing data for discovering valuable information in large data sets with knowledge of statistics, machine learning and database system.
Pattern recognition	A branch of machine learning that focuses on the regularities in data
Deep learning	A branch of machine learning based on complex structure of neural networks
Artificial intelligence	The study of intelligent systems and agents with the ability of learning from circumstances and solving problems

From a general point of view, the data analytics or data mining is the computational process to reveal the potential relations between variables with the techniques including database, statistics, pattern recognition, machine learning, etc. However, due to the diverse sources, the collected data sets may have different levels of quality in terms of noise, redundancy and consistency.

1.2.1 Data Pre-processing

The data pre-processing techniques are necessary to improve data quality as shown in Figure 1.3. Data integration techniques aim to aggregate data collected from disparate sources in an effective way with a unified view [25]. For example, when combining the datasets of weather condition records and power system interruption events, the attribute of “date time” would appear twice. But apparently only one attribute of “date time” is needed for the following data analytics process. The same attributes with different name as well as the different attributes with the same name is to be identified in this process [26]. Normally, the correlation analysis is used in the redundancy identification to abandon the highly correlated attributes and reduce the size of datasets. In most cases, the datasets would contain some missing values which influence the results of data analytics. Deletion or interpolation are the frequent techniques to solve such kind of problems. As to the abnormal values, the first step is to check whether this is rational based on the professional knowledge for the application. If it is caused by

an error in sensors or data processing platform, we can treat it as a missing value or try to find the real value, otherwise it is supposed to be kept in the dataset as a "black swan". The logarithm is an effective way to "correct" the distribution shape of data with severe skewness, because some data analytics algorithms are sensitive to imbalanced data. New attributes such as the temperature difference can be calculated in the pre-processing step if there is only maximum and minimum value of temperature in the initial dataset. The new constructed attributes are usually helpful to improve the accuracy of data analytics results.

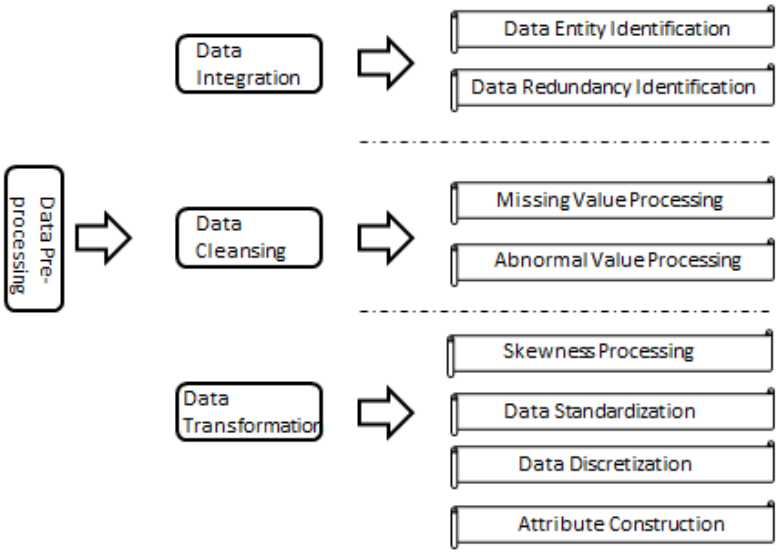


Figure 1.3 Data Pre-processing Techniques

1.2.2 Data Pre-processing

The most frequently used data mining or machine learning algorithms are usually categorized as supervised or unsupervised learning depending on whether there is a label attached to each item in datasets as shown in Table 1-5. For the supervised learning algorithms, the data analytics model can be trained based on the given data to discover the relation between data attributes and the corresponding categories or values. While for those without labels, the data analytics model is usually designed to recognize the possible groups among all the items [27].

Table 1-5 Data Analytics Algorithms

Category	Algorithm	description
Supervised Learning	Decision tree	A non-parametric method with a tree-like method whose leaves represent class labels and branches represent conjunctions of features
	Naive Bayes	A probabilistic method based on Bayes theorem with the assumption of independence between every pair of features
	Support vector	An algorithm to find a separating

	machine classifier	hyperplane between the two classes by mapping the labelled data to a high-dimensional feature space
	K Nearest Neighbor	A non-parametric method based on the minimum dissimilarity between new items and the labelled items in different classes
	Random Forest	An algorithm consisting of a collection of simple tree predictors independently for the estimation of the final outcome
Unsupervised Learning	K-means	An unsupervised learning method with a given number of clusters to sort the data based on the average value of data in each group as the centroid
	K-medoids	An unsupervised learning method similar to k-means by assigning the centroid of each group with an existing data point instead of the average value
	Hierarchical Clustering	An alternative approach which aims to build a hierarchy of clusters in a dendrogram without a given number of clusters
	DBSCAN	A density-based clustering algorithm to identify clusters with specific shape in distribution
	Expectation-Maximization	An iterative way to approximate the maximum likelihood estimates for model parameters
Correlation	FP-Growth Algorithm	An efficient method for mining the complete set of frequent patterns with a special data structure named frequent-pattern tree with all the association information reserved
	Apriori Algorithm	A classical data analytics algorithm to discover the potential association rules among frequent items
Dimensionality reduction	Principal Component Analysis	An orthogonal transformation of data with a new coordinate system with the greatest variance projected to the first coordinate
	Self-organizing Map	A type of artificial neural network for a low-dimensional representation of the training data space
	Random Matrix	An algorithm which reveal potential regulations with high order matrices for massive data by eigenvalue analysis

1.2.3 Procedures of Data Mining in Smart Grids

As shown in Figure 1.4, the main procedure of data analytics in smart grid is to extract valuable information from historical data for guiding the operation and maintenance with the comparison to real-time data [28]. The huge amount of data

collected from smart meters and sensors are arranged and stored with data management techniques. After preparation, the mathematical model can be established through data mining techniques based on the clean data. With the input of real-time measurements, the state status can be evaluated in the derived model, which provides the possible schemes to guide practical actions and solve potential problems.

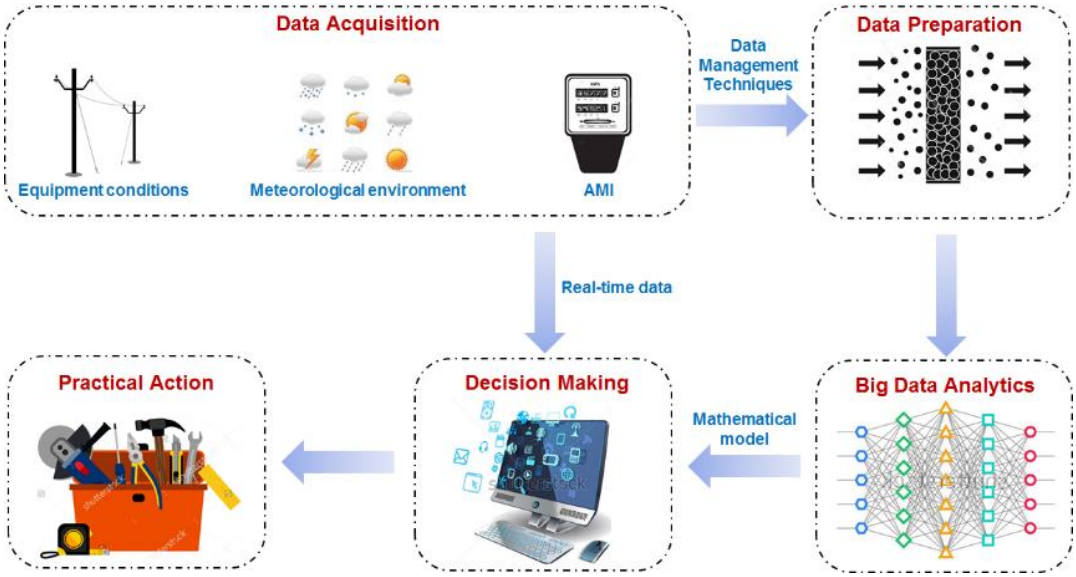


Figure 1.4 Example of big data analytic procedures in smart grid

1.3 Big data analysis in smart grid

1.3.1 Fault detection

The carbon emission reduction and sustainability of environment are the driving force and construction purpose of smart grid, which is designed in a decentralized structure. The employment of distributed generator units in modern power distribution system now provides an effective means for the utilization of widespread renewable energy such as wind and solar energy. These emerging microgrids are vital for the expectation of a low-carbon society. Moreover, the close distance between the generator and loads in microgrid improves the reliability of power delivery and reduces the power transmission loss. The ability to operate in an island mode also protects the load from damages caused by power system including voltage fluctuation, frequency deviation, etc. [29].

However, the intermittent characteristic of renewable energy increases the uncertainty in power grid, whose typical solution is to use inverter interfaced distributed generators (IIDGs) for a better power quality. In contrast with the traditional bulk generators like large-volume thermal, nuclear or hydro generators, the much lower inertia of IIDGs is a severe potential threat when the faults in microgrids cannot be detected and cleared in a short time due to the limited current carrying capacity. Most of the traditional techniques relying on the detection of overcurrent and negative sequence current origin from the large-scale centralized power system and seem less effective in microgrids. A statistical

classifier-based protection scheme using local current measurements is proposed by applying the wavelet transform and the decision tree (DT) model in [29]. The wavelet transform can decompose the signal in time-frequency domain with the time localization reserved. Energy, Shannon entropy and standard deviation of the wavelet coefficients which contain the information during transient events are calculated. Finally, 15 statistical features extracted from the current data for one cycle by sequence analyzer and wavelet transformation are fed into the DT models for fault detection and classification. A differential protection scheme for microgrid is proposed in [30] with the most sensitive features at both ends of the respective feeder processed by the discrete Fourier transform. These differential features are then utilized in the decision tree-based data-mining model for determining the final relaying decision.

For a grid-connected microgrid, the severe weather conditions or grid blackouts may trigger an unintentional islanding accident, which threats the safety operation and causes technical issues. Artificial neural networks (ANNs) are trained in [31] with features extracted from the differential transient of the rate of change of frequency (ROCOF) signal in order to identify islanding accidents. A support vector machine (SVM) classifier is established in [32] with multiple features extracted from system variables as an islanding detection approach. The feature extraction process is implemented with a sliding window whose width is optimized for the highest detection rate.

As a real-time social sensor for the smart grid, social media like Twitter or Facebook could contain potential information indicating the occurrence and location of power outages [33]. A probabilistic framework is devised in [34] for detecting a targeted event from the fragmented and noisy tweets. The method shows a good performance in locating accrual outage areas in experiment, which could be integrated to a social data-driven outage management.

1.3.2 Predictive maintenance/Condition-based maintenance

Distribution automation (DA) is a concept of smart grid which focuses on the operation and system reliability at the distribution level. A successful DA has the capability to localize and isolate the faults in distribution system with a reduced restoration time and improved customer satisfaction. Under the concept of DA, increasing volume of operational data have been collected from supervisory control and data acquisition (SCADA) or advanced metering infrastructure (AMI) for state monitoring and fault diagnosis.

Reference [35] proposes an analyzing scheme for preventative measures to avoid or minimize the outages with the data related to pole mounted auto-recloser (PMAR). PMAR is a kind of protection intelligent electronic device installed on the overhead lines of a distribution network which attempts several recloses after an interruption happened in the downstream of the feeder.

Thanks to the development of ICT technology in power systems, a huge volume of data can be collected via AMI and communication infrastructures. Power system operating data, weather information and log data of relay protection

devices are processed as the input of a one class classification system, which is a data-driven model of fault phenomena based on a hybridization of evolutionary learning and clustering techniques in [36], [37]. This fault recognition system is validated in the medium voltage power grid in Rome. The traditional statistical methods such as linear discriminant analysis (LDA) and logistic regression are discussed for mining the relation between power system faults and the features extracted from raw data [38].

As a potential threat to the security of transmission systems, the galloping of power lines can cause structural and electrical failures. After analyzing the impact factors of galloping, a data-driven model based on SVM and AdaBoost bi-level classifiers is proposed in [39] for early warning. The extreme learning machine (ELM) algorithm is applied in an intelligent early-warning system for reliable online detection of risky events in power system in [40]. Since the weights in ELM training are randomly chosen and then determined through matrix computation without iterative parameter adjustment, the learning speed is much faster than conventional algorithms, which is an ideal solution in "big data" cases. The optimal balance between earning accuracy and warning earliness of the data-driven framework is also discussed. Reference [41] provides a method to extract electrical features from high-impedance fault current and voltage signals and build an effective feature set (EFS) via a ranking algorithm. Therefore, only a small number of signal channels are required to build a statistical classifier for fault detection. Reference [42] also provides an effective method to reduce the huge volume of PMU data while retaining the critical information for fault detection in power system.

1.3.3 Transient stability analysis

Transient stability is a critical issue closely related to the safely operation of power system. However, the increasing demand for electricity, growing penetration of renewable energy sources and deregulated market force power grid to operate near their secure operating limits [43]. Facing with the challenges from a more complex system, transient stability analysis (TSA) for the study of dynamic behavior taking the electromechanical and electromagnetic process in power system taken into consideration is becoming a hot research topic. The transient process and new operating conditions need be calculated with the TSA technique after a severe interruption in power grid for a comprehensive protection scheme. Traditional TSA based on the time-domain simulation is not able to provide universal results due to so many uncertainties.

Under the concept of smart grid, a large amount of data collected via AMI are involved in the state assessment of power systems to support the energy management, system operation and decision making. Therefore, efficient summarization techniques are required for extracting useful patterns and discovering valuable information from redundant measurements in power system. A DT-based framework is proposed in [43] [44] for the dynamic security assessment (DSA) in power system with high penetration of DGs. Two

contingency-oriented DTs are trained based on the databases generated from real-time simulations. One of the well-trained DT is fed with real-time wide-area measurements to identify potential security issues, and the other DT provides the online corresponding preventive control strategies to deal with the problems. In [45] the dominant instability generation group (DIGG) in power system is identified without time domain simulation since the features adopted for TSA are extracted from steady-state variables. Reference [46] proposed an approach to classify the collected data from smart grid into two classes called vulnerable and non-vulnerable data sets with the data analytics such as multichannel singular spectrum analysis (MSSA), principal component analysis (PCA) and SVM. A framework for online contingency screening is presented in [47] with respect to first swing transient stability. The large spectrum of pre-fault operating state variables and critical clearing times of several contingencies are collected to compose a dataset for pattern recognition methods. The metric which can be used for operating condition evaluation is developed through PCA.

In addition to the renewable energy micro-sources distributed in smart grid (SG), the grid-connected high capacity wind farms are also widely accepted and applied for an effective utilization of pollution free and abundant nature resources. The improvement of technologies for large wind turbine generators and high capacity power converters accelerate the amount of wind energy integration into power system. To address the potential deterioration and stability problem caused by the large integration of wind energy to power grid, reference [48] proposed data-driven analytics to determine the Q-V characteristic curve at the point of interconnection of the wind farm with valuable information for voltage stability extracted. Without prior knowledge of the system configuration and parameters, different curve-fitting techniques are adopted in a real case study in Canada.

Power swing is the oscillation of power flow on transmission lines when the angles of rotors of synchronous machines are advancing or retracting to each other which may cause a large disturbance. Heavy load shedding, generator triggering and short-circuit faults clearance are all the potential reasons. Reference [49] used a decision tree-based scheme for fault detection and classification during power swing within half cycle time. The decision tree algorithm is also adopted in [50] with 21 potential features extracted from phasor measurement unit (PMU) data after Kalman filter process for intelligent relaying in transmission system. A probabilistic framework is established in [51] based on the decision tree and hierarchical clustering for dynamic behavior of power systems after an occurrence of interruption. The unstable groups which may lose synchronism can be successfully detected.

Although the PMU and WAMS provide high-resolution datasets for engineers to discover patterns of normal and abnormal operation, the low probability of events that occur in power grid leads to a severe class imbalance problem. The conventional data analytics are difficult to extract the features of rare instability from massive synchrophasor measurements. Reference [52] develops a systematic imbalance learning machine for online short-term voltage assessment. A forecasting-based nonlinear synthetic minority oversampling technique is adopted

in the cost-sensitive learning algorithm to deal with the class skewness. To take full advantage of massive power grid data, the random matrix theory is introduced in [53][54] with a high-order data-driven model to present the power system parameters and external data like meteorological information. The eigenvalue-based analysis method is proven to deal with online transient state analysis. An online monitor of instantaneous electromechanical dynamics in transmission system is presented in [55] based on the parallel computing and k-nearest neighbors learning algorithms. The information that indicating time-varying correlations of power generation and consumption is extracted with the proposed framework. An active learning solution is proposed in [56] to solve the problems for online data-driven model updating and offline training, which provide an efficient way for data sets preparation. A novel PMU-based robust state estimation method is proposed in [57] for online state estimation of a power system under different operation conditions with the help of an adaptive weight assignment function to dynamically adjust the measurement weight according to the large disturbance revealed from PMU data. A similar framework is proposed in [58] to enable the utility company for real-time data processing. The core vector machine (CVM) is used for a two-class classification in [59] to process the huge amount of PMU data from power grid. The CVM model is trained offline with 24 features extracted from the raw data for an online assessment evaluation for the TSA problem. The transient stability boundary of large-scale power systems is analyzed in [60] by a statistical nonparametric regression methodology based on the critical clearing time to determine whether a steady-state condition can recover after a given fault.

1.3.4 Electric device state estimation/Health monitoring

As a vital component for electrical energy conversion, a failure in power transformers may cause catastrophic blackouts in power system [61]. Therefore, the life-cycle management of power transformers based on an accurate estimation attracts a lot of researches for a more stable and reliable power grid. The existing diagnosis methods for power transformers mainly focus on limited state parameters with the threshold-based diagnosis. To take information of system operation and meteorological conditions into state estimation analysis, three classical algorithms for association rule mining are discussed in [62], namely, Apriori, AprioriTid and AprioriHybrid. The rule mining methods are combined with probabilistic graphical model for potential failure prediction.

In most commercial buildings, the building automation system (BAS) are designed and adopted to control the heating, ventilating and airing conditioning (HVAC) system to maintain proper temperature and humidity for the occupants. If the indoor smart grids can be monitored on a continuous or regular basis, a proper operation strategy may be proposed for the improvement of energy efficiency, fault diagnosis and system reliability. In [63] a novel health monitoring system is proposed by the fuzzy logic for abnormal operating condition detection. The fault

signatures for various fault types are generated by the ANN classification technique.

As the rising number of aging assets in power system is becoming a potential threat to the safety operation, a lot of failure models are proposed focusing on variables of aging time or conditions. Reference [64] proposed a failure rate model for general electric power equipment with the lifecycle data of service age, maintainer, health index taken into consideration. In order to make the best use of these data, the stratified proportional hazards model (PHM) is developed as a nonparametric regression method to process and classify the lifecycle data into multi-type recurrent events quantitatively [65]. The potential risk problem and health condition can be predicted with the help of this PHM method [66].

1.3.5 Power quality monitoring

As a worldwide issue, Electric power quality (PQ) refers to the magnitude, frequency and waveform of voltage and current in power system and highly related to the safe operation of power grid as well as the satisfaction of consumers. With the increasing application of nonlinear and power electronics based loads and generators, the harmonic distortions and instable situations frequently appears in power grid. Deep learning is successfully employed for the classification of PQ events of the electricity networks in [67]. Instead of sampling the voltage data of the PQ event data like the existing analysis methods, the image files of the three-phase PQ events are processed for classification by deep learning techniques. Due to the high cost for installation of advance metering devices, the conventional electromechanical analog meters still work in some residential areas and the data analytics-based PQ analysis cannot be properly utilized. Reference [68] presents a framework that collecting electricity information of from analog meters via image processing techniques. The power consumption information can then be collected to a cloud server through online data exchange. Under the consideration of balance between computation capability and the satisfactory performance of the algorithm, a compact method is presented in [69] for feature extraction from the raw data in smart grid to get information that is highly related to the field of power quality. A robust and fast processing pattern recognition algorithm is proposed in power quality events (PQE) classification is illustrated in [70]. The features highly correlated to the PQE are extracted with the discrete wavelet transform-entropy and basic statistical criteria for the establishment of ELM classifier.

1.3.6 Topology identification

Taking the advantage of information layers in smart grid is an effective means to approach the challenges from the renewable energy sources (RES) in distribution network. The measurement, monitoring, communication and control of smart grids by advanced sensors and devices are making the complex network sensible and perceptible. The randomness of RES and uncertainty of the load are

increasing the urgency and necessity for a comprehensive decision based on huge volume of data collecting and processing. The SCADA and WAMS provide voltage and power data of smart grid in near real-time sampling rate [71] [72]. Since the network-constrained economic dispatch problems are supposed to be solved by the real-time electricity process in a contemporary whole-sale electricity market, the potential of recovering the topology of a grid is explored with market data in [73]. Another dynamic solution for online SG topology identification (TI) is proposed in [74] which is reformulated as a sparse-recovery problem. Grapy theory and probabilistic DC optimal power flow are adopted for building the network model.

With the purpose for a greener society, the low carbon technologies (LCTs) are driven by the government by application of heat pumps, photovoltaic, electric vehicles and other smart appliances in low voltage (LV) distribution networks. Therefore, the visualization of LV networks with limited metering and data acquisition equipment attracts increasing research interests. The network load profiling based on the identification of representative load profiles of LV systems is an economical alternative method. A novel three-stage network load profiling method proposed in [75][76] aims to evaluate the capabilities of the current LV networks to accommodate the LCTs by clustering, classification and scaling. The first two stages are used to identify the load conditions of unmonitored LV systems with similar fixed data to those monitored LV substations. The contribution factor for each LV template is then determined by the cluster-wise weighted constrained regression algorithm.

1.3.7 Renewable energy forecasting

The abundant and environmental friendly RES such as wind and photovoltaic energies are supposed to be the dominant energy source for the next generation of power grid. However, the randomness and intermittent characteristics are always obstacles for a large-scale utilization of RES in a stable way. To deal with such enormous challenges and get an improved dispatch planning, maintenance scheduling as well as regulation, an accurate and reliable RES forecasting approach has become the hot spot around the world [5]. A data mining based method consisting of k-means and neural networks is proposed in [6]. The meteorological information in historical records are used for clustering approach to classify the days into different categories. Then the bagging algorithm based neural network is trained to get the forecasting results of wind energy. Instead of using the neural network, [7] utilizes the support vector regression method to predict the wind speed with the time series historical wind speed processed by empirical mode decomposition into several intrinsic mode functions and residue. In [77] a short-term probabilistic wind generation forecast method is presented based on the sparse Bayesian classification and Dempster-Shafer theory as a nonparametric approach. Reference [78] studies the ultra-short-term wind forecasting with the deep learning method through unsupervised feature learning from the unlabeled historical wind speed data. The forecasting approach of

distributed solar energy systems from macro- and micro-aspects is discussed in a general way in [79] with clustering of capacity and location of PV system. The data-driven forecasting approach of PV diffusion is proposed based on cellular automation in microscopic analysis. By decomposing the time-series data with discrete wavelet transform, the proposed recurrent neural network (RNN) model in [80] is developed for ultra-short-term solar power prediction.

1.3.8 Load forecasting

Like the RES prediction, an accurate short-term load forecasting is the essential basis for energy management, system operation and market analysis. As is mentioned in [81][82], an increase of forecasting accuracy may bring a lot of benefits and save the investments. With the emerging active role of customers in smart grid, the high efficient dynamic electricity market is also based on a good performance of electricity consumption prediction. Since electricity consumption is affected by the weather conditions to some extent, reference [83] proposed a Map/Reduce programming framework for distributed load forecasting by partitioning the geographical area according to local weather information. An extreme learning machine ensembled with a novel wavelet transform is used for electricity consumption in [84] after a conditional mutual information based feature selection, which is also used in [85]. To overcome the volatility and uncertainty of load profiles, the recurrent neural network is adopted with a novel pooling layer to avoid overfitting problems in [86]. Rather than the aggregated load forecasting, the energy consumption in a single house is usually volatile and difficult to be predicted. Driven by the recent success of deep learning, a long short-term memory recurrent neural network based framework in [87] is applied to the residential load forecasting as the latest deep learning techniques. A hidden mode Markov decision process model is developed in [88] to the forecast the customers' real-time behavior. Reference [89] analysis the emerging trends and challenges in the new era of using social media through mobile apps to improve their customer engagement and load forecast. Reference [90] considers the impact of social activities on the prosumers' arrangements for their generation and consumption patterns and further discuss the overall impact on the final load and the network usage.

1.3.9 Load profiling

Load profiling is a way to describe the typical behavior of electric consumption, which is usually represented in time domain for load forecasting, demand-side management and capital planning [91-95]. As an effective method for energy management, the tariff structure designed before is usually based on the type of activity, which is not able to indicate the electrical behavior in a comprehensive way [96]. Reference [97] utilized a two-stage clustering algorithm to classify customers according to their load curves. In the first-stage, the load patterns are clustered into different categories according to the evaluation index,

and then the customers are classified according to the comprehensive load shape factors defined in the first-stage with SVM algorithm. In contrary to the time domain analysis [98], the DFT method is adopted in [99] to discover the information of customers' behavior, which can be accurately reconstructed using limited frequency components and still satisfy the strict requirements. The residential electricity consumption usually can be divided into three parts: fixed, regulable and deferrable loads, which is the theoretical basis for the optimal energy management of the demand response (DR) mechanism. DR is used to initiate a change in the customers' consumption or feed-in pattern with an incentive from costs or ecological information. Reference [100] utilizes the spectral domain analysis methods DWT and DFT to decompose smart metering data with the extracted coefficients. Results show that DWT performs better than DFT in individual level while DFT is more suitable to be used in the analysis at a highly aggregated level. A learning-based DR strategy combining data analytics and optimization is developed for regulatable loads focusing on the residential HVAC [101]. Because when the customers' behavior is obtained, an optimal DR technique for household HVAC unit can be designed based on weather prediction, day-ahead electricity price. Reference [102] takes the advantage of the social networking to minimize the peak power consumption of the electrical appliances by proposing a "family plan" approach which leverages the social network topology and statistical energy usage patterns of the users.

To better understand the information behind the stochasticity and irregularity of residential energy consumption, an in-depth analysis is presented in [103] with a finite mixture model-based clustering technique. The self-organizing maps (SOM) as a type of ANN is used in [104] to reduce the dimension of collected raw data for load pattern extraction. The frequency-domain data analytics in the SOM shows a superiority over the time-domain data with a higher accuracy in new customer classification. As one of the main tasks of load profiling, a better understanding of the flexibility of customers' electricity consumption is the basis for DR, which can be used to release the pressure of distribution system in terms of thermal and voltage constrains. A multi-resolution analysis method based on wavelet analysis is proposed in [105] to extract spectral and time-domain features of load data. Different permutations of typical load profiles provide a more flexible load profiling with a reduction of computation. With the popularization of electric vehicles (EVs), learning the charging load patterns of them is becoming a key step for the stability of power grids. An unsupervised clustering algorithm is used in [106] to extract the pattern of EV charging loads with only the real power measurements. Furthermore, the flexibility of the collective EV charging demand is analysed with Bayesian maximum likelihood. References [107] and [108] focus on the problem brought by the huge load profile data with the popularity of smart meters installed at the household level, which poses challenges to the communication and storage of measurement data as well as the vital information extraction from massive records. K-SVD sparse representation technique is used to decompose the load profiles into several partial usage patterns for a linear SVM based method to recognize the type of customers.

1.3.10 Load disaggregation

Load disaggregation is also called non-intrusive load monitoring (NILM), aiming to segregate the overall load profiles at household level into the energy consumption of individual appliances. Unlike direct appliance monitoring framework, the NILM from only one smart meter installed in the house is easier to be accepted by the customers [109][110][111]. Since different types of the household electric appliances have different potential to be involved in the DR program, the appliance-level load profiles allow the utilities to understand the customers' behavior better and helps to develop a more energy efficient strategy. The early techniques for NILM are mainly based on the detection of "edge" in power signal to indicate the state "on" or "off" of a known device [112]. The more effective and complex appliance signatures are then proposed with the harmonics computation of steady-state power or current [113][114]. The hidden Markov models (HMMs) are adopted in [115] with the segmented integer quadratic constraint programming to disaggregate the household power profile at an average frequency of 0.3 Hz into the appliance level. In [116] a NILM approach based on the subtractive clustering the maximum likelihood classifier is proposed for a low-sampling-rate date set of 1 Hz sampling rate. The appliances are modeled as ON/OFF states in this event-based load disaggregation algorithm. As a single channel blind source separation problem, the dictionary learning based approaches can be used in NILM. A deep learning approach with multiple layers of dictionaries trained for each device as "deep sparse coding" is utilized in [117][118]. Compared with HMM, the latter method is not suitable for real-time application. By combining the decision tree and nearest-neighbor algorithms, the semi-supervised machine learning is applied to the NILM problem in [119] with the signal features extracted by matching a set of net wavelets to the load classes.

1.3.11 Non-technical loss detection

The nontechnical loss (NTL), which is probably caused by the electrical theft or errors in accounting, is one of the prominent concerns that have plagued the power system utilities for a long time [120][121][122][123]. According to the survey published by Northeast Group, LLC, the loss caused by electricity theft reached more than $89.3 billion in the world every year [124]. Furthermore, large scale electricity fraudulent behavior may cause severe imbalance problems in power system. Therefore, the effective framework to detect the NTL in the complex power grid has appealed many research interests. A comprehensive top-down scheme based on DT and SVM is proposed in [102]. DT is trained with various features including heavy appliances, number of persons, weather conditions to get the expected value of electricity consumption for the customer during a particular time. Then the calculated consumption along with other features are fed to the SVM classifier which is already trained based on the collected dataset to determine whether the customer's behavior is normal or fraud. In [125] the fraud detection is triggered when a discrepancy is detected between

energy supplied from the power system and collected information from smart meters. The anomalies in consumption patterns are discovered with the fuzzy clustering algorithm.

1.3.12 Open issues for the application of big data analytics in smart grids

Even though there are increasing researches on the big data analytics in smart grids, the deployed applications are few. There are still many open issues needed to be addressed before the techniques can create implications in reality.

With the fast deployment of smart meters and advanced sensors, huge amount of data with multiple types and structures from deference sources with a variety of protocols are generated every second. However, the lack of standard data format for the information software and database structures, as well as the issue of interoperability of different information and communication systems deployed in the smart grids, make it complicated and difficult to obtain data for real application. The traditional way of isolated storage of the data in various systems also increases the barrier for data sharing among applications.

As a conventionally sensitive industry, most of the data generated in the smart grid are considered as confidential or related with privacy issues; therefore, it is impractical for researchers to conduct highly relevant studies which can be smoothly transferred later on into deployment. Thus, most of the researches are still about the algorithms which are tested with ideal data, and hence stay in the Ivory tower.

In addition, due to the lack of strategic vision, top design of application, large investment in reality, combined with the short-sighted recognition of the value of the data, the applications of big data in real systems are growing very slow. Even though, the majority of utility companies showed great interests in the big data analytics and their application in their business, they are still waiting to see convincing results before they are willing to put more efforts and investment.

Last but not least, the big data analytics in smart grids is a comprehensive and complicated field, which does not only depend on the mathematic algorithms or techniques, it also depends on the operation of the systems, the behaviors of vast number of autonomous users, the ICT technologies, the expertise of the field, etc. Therefore, it needs the synergy among experts from different fields if we would like to see the benefits of it in the smart grids.

1.3 Summary

In this article, the big data in smart grid and the corresponding state-of-the-art analysis methods have been reviewed and discussed. The data which may contain valuable information are collected from smart meters installed in the power system, electricity market, GIS, meteorological information system, social media, and so on. The purpose of advanced ICT technology in power system is to associate the traditional physical parameters in power system to the external

variables to discover potential regulations and scientific problems. Eleven applications of data analytics mentioned in the paper are nearly involved in every aspect of smart grids, including the operation, maintenance, load/output forecasting, protection as well as fault detection and location. After extracting the useful features from raw information with the background knowledge of electrical engineering, typical data analytics methods, such as neural network, k-means, and support vector machine, could be widely applied. Secure and efficient operation strategies as well as optimal business decisions are supposed to be made with the data analytics from a more unified view.

With more advanced ICT technologies applied in power system, the fast and efficient data analytics framework for huge volume of data would become a challenging requirement. Moreover, the cyber security and privacy protection could become as important as a relay protection in power system. Even though the interactive communication with customers provides a potential solution for more accurate demand response, it also increases the difficulties in consumption behavior analysis at the same time. A secure and high-performance data analytics platform would be crucial for the social welfare and power companies' interests in the future. As the application of data analytics in smart grids is a comprehensive and complicated field, involving mathematics, ICT technologies, computer science, electrical engineering, etc., thus, it needs the synergy among experts from different fields as well as the strategic visions for the top designs.

1.4 Structure of the following content

In order to identify feeders with typical and representative characteristics in the distribution network, the clustering algorithm is used in Chapter 2 to classify and analyze the structural characteristics over more than 9,000 cases. It can be found that different types of feeder structure have a significant effect on the probability of fault. This provides a scientific basis for local grid companies to keep track of distribution network security operations and for predictive maintenance.

Chapters 3 and 4 employ the similar data models, i.e. supervised machine learning algorithms, to enable the utilization and analysis of two completely different kinds of data in the power system. Chapter 3 focuses on the prediction of faults and can be divided into two parts: the prediction of the number of faults and the prediction of the severity of faults. The former, in years or months with a large time span, uses the prediction method of grey theory to obtain the number of future faults in the local distribution grid over a longer period of time. The latter predicts, in days, the likely occurrence of a fault in the coming day or days in conjunction with the weather forecast. This provides a certain reference for maintenance arrangements for the distribution network. Chapter 4 uses a random forest algorithm, to perform the regression analysis of users' electricity consumption combined with weather records in every 15 minutes. By measuring the degree of correlation between the electricity consumption and corresponding weather conditions, it is possible to obtain the degree of dependence of customers for different weather features and use it as a basis for detecting abnormal users.

The last chapter discussed the effect of heat waves on the local distribution network, which is a high-impact low-probability event considering the repetitive faults within a certain time period. The probability of heat waves and the probability of repetitive faults

under such conditions are calculated based on the historical records, based on which the Monte Carlo simulation could be built with Poisson process. The cost and benefit of an investment are analyzed as an expectation over tens of thousands of simulations. This could be an important reference for the power system planning in practice.

Chapter 2

Identification of representative MV Feeders in distribution network

2.1 Introduction

With the application of distributed renewable energy and electric vehicles, the structure of medium voltage (MV) feeders becomes increasingly complicated. While as the terminal of power grid, distribution feeders access directly to users and have an important role in the quality of power supply. Due to the different profiles of customers and purpose for infrastructure planning, the distribution lines vary widely in their structural features, including the number of customers, capacity and neutral grounding modes. However, there are limited number of sensors in the distribution network and the sampling interval of smart meters is usually too large for the algorithms to give a rational real-time analysis result. With tens of thousands of MV lines distributed in an area, it also takes too much time and effort for distribution network operators (DNOs) to have a detailed analysis for each unique feeder [126-127]. Therefore, the performance evaluation of MV feeders in distribution network has attracted researchers' attentions for reliability analysis of distribution networks.

Data mining technology is an efficient analytic tool to sort the mixed data in a rational way and extract representative information from complicated data sources, which has already been successfully implemented in distribution power grid [128-129]. Clustering algorithm is a good method for the researchers to find out the taxonomy of prototypical feeders based on the statistical analysis. In north America, 24 radial distribution feeder models are presented as the typical results from total 575 distribution feeders [130]. The significant parameters from engineering views are selected from initial set of parameters for clustering in [131], which reduced the computational complexity and improved the quality of clustering results. Based on the common structure features extracted from the plenty of MV feeders, it is also possible for engineers to choose the best technical

and economic planning of smart grid. Clustering analysis is utilized to identify the representative high voltage feeders in the West Australia in [132] and the results are taken as the input of models to assess the impact of technology changes in smart grid. In [133] the k means algorithm is adopted to group the classify distribution feeders into specific groups.

2.2 Features of feeders in distribution network of Italy

2.2.1 Feature description

Two sets of MV feeders are collected from an electrical company in Italy, among which there are over 9000 feeders (15kV) distributed in the north are of Italy and over 10,000 feeders (20kV) mainly in the middle and south part of Italy. The number of feeders in each region of Italy is shown in Figure 2.1. Every feeder contains 11 structural features for the clustering process as shown in Table 2-1.

Table 2-1 Structural features for each feeder

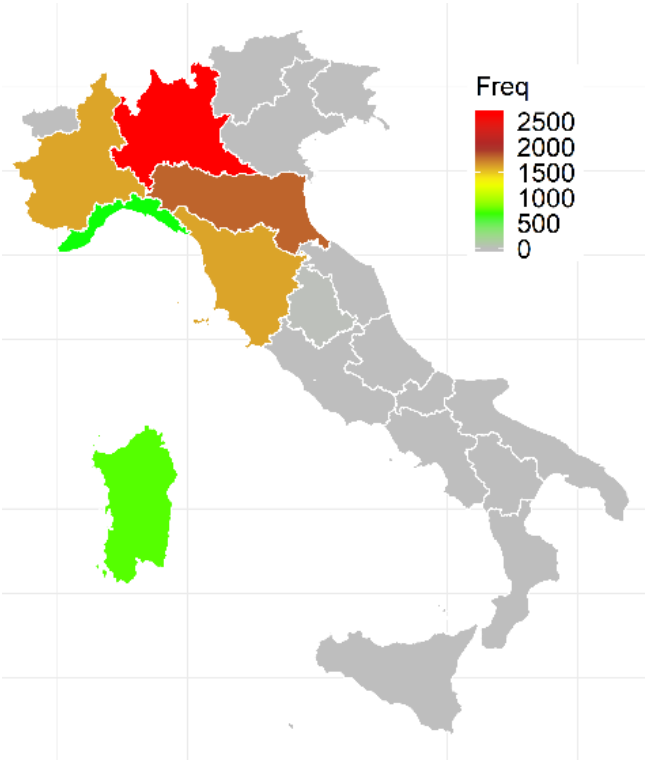
Feature	Description
Length	Total length of the feeder
Cable%	Percentage of underground cable in a feeder
Nodes	Number of nodes in a feeder
Branches	Number of branches in a feeder
Customers	Number of customers in a feeder
Sec Sub	Number of secondary substations in a feeder
Auto Nodes	Number of nodes with automation equipment
MV/LV Trans	Number of MV/LV transformers in a feeder
Capacity	Apparent power in a feeder
Neutral Types	Neutral grounding mode of a feeder
Auto Types	Automation types of a feeder

The first 9 features in Table 2-1 are numeric data corresponding to the electrical characteristics and power delivery capacity of each feeder. For example, the percentage of underground cables has a profound impact on the impedance of power lines and the total length of each line has a relationship with the resistance and affects the probability of interruptions to some extent. Generally, the more number of customers a feeder serves, the more likely it would suffer an interruption. The nodes with automation equipment means that there are remote controllers on crucial nodes in the middle of feeder, which would isolate fault parts resupply the customers after transient interruptions.

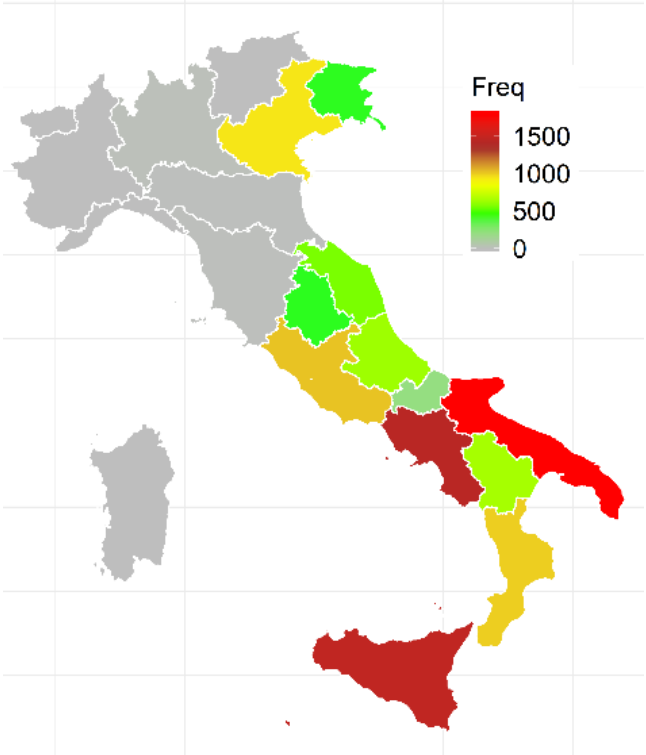
The last two features are categorical data including neutral grounding modes and automation types of each feeder as shown in Table 2-2.

Table 2-2 Neutral Grounding Modes and Automation Types of 15kV feeders

Neutral Grounding Mode	Isolated	Resistance	Fixed Coils	Adjustable Coils	Fixed+ Adjustable Coils	
Automation Types	FNC	FRG	FNC+ICS	FRG+ICS	ICS	None



(a) Distribution of 15kV feeders in Italy



(b) Distribution of 20kV feeders in Italy

Figure 2.1 Distribution of MV feeders in Italy

For the mode of neutral grounding in MV feeders, there are five different types taken into account. These neutral grounding types have different impact on the performance of the distribution network. For example, when there is a single-phase grounding failure, the fault current is equal to capacity current for isolated neutral point. But if this capacity current can be well compensated by the coils in the neutral point, the fault current is limited to a relatively small value so that the failure can extinguish itself in most cases. The interruption records for performance evaluation are those whose fault duration persists more than one second. Transient interruptions are not taken into consideration.

As for the automation type, two main schemes FRG and FNC are widely adopted. The former one can be used for high fault current because the circuit breaker in the primary substation of a distribution feeder trips immediately when the fault is detected. However, it might take much time for switches in automated nodes to do reclosure. According to the standardization of Enel distribution company, the total time for FRG protection is limited to 180s. Moreover, the healthy section of the feeder may suffer several short interruptions before isolating the fault part. The latter automation type is supposed to be applied on the situation where fault current is lower enough for switches to open according to the instructions from the remote terminal unit (RTU). Therefore, the switch can be used as a circuit breaker and isolates the fault section from far to near the primary substation without any interruptions on the healthy parts of the feeder. The automation of ICS means a second circuit breaker installed in the middle of the feeder on the secondary substation, which coordinates with the one installed in the primary substation to isolate the fault section. The only ICS automation is pretty rare since it usually cooperates with FRG or FNC.

2.2.2 Feature selection

Although there are 11 features of each MV feeder from the dataset, some of the features are redundant because they are highly correlated, which could behave in a very similar manner in data analysis and will make computation much time-consuming [134]. In the following analysis, we will focus on the 15kV feeders' analysis. One intuitive way to find the redundant features is the scatter plot of all the numeric features are shown in Figure 2.2.

As shown in Figure 2.2, the number of nodes in a feeder has a strong correlation with its number of branches, which reveals the fact that in most cases there is only one branch out of the node in a feeder. The similar circumstance happens between the secondary substations and MV/LV transformers. In statistics, the relevance between two variables can be quantitatively described by the Pearson correlation coefficient [135] as shown in Equation (2.1).

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where the cov(X,Y) is the covariance between variable X and Y, σX and σY are the standard deviations of X and Y, respectively [136-137]. The value of Pearson

correlation coefficient is between -1 and 1. The features with absolute value of Pearson correlation coefficients larger than 0.9 indicate a very strong dependency between each other. When the value is 0, it means two features are totally independent from a statistical point of view. Positive value means positive correlation and vice versa. The Pearson correlation coefficients of numeric features are shown in Figure 2.3, according to which the number of nodes, secondary substations and branches are to be deleted.

For the remaining categorical features, it is easy to plot the number of feeders equipped with different combinations of neutral grounding modes and automation types. As can be seen from Figure 2.4, there is no feeder with an isolated neutral point and FNC or FNC+ICS automation type in the same time. This is because both two automation schemes have a long-time delay before circuit breaker trips, which can only be applied only when fault current can be limited to a relatively small value. The coil grounding mode in neutral point of distribution system can compensate the capacitive current through neutral wire and guarantee the fault current to a limited value. Figure 2.4 also shows that there are few feeders in distribution network with a resistance grounding system or ICS automation equipment, which account for only 1% of the total feeders. Therefore, these two categorical features are better to be eliminated to avoid the adverse effect for the following data mining method.

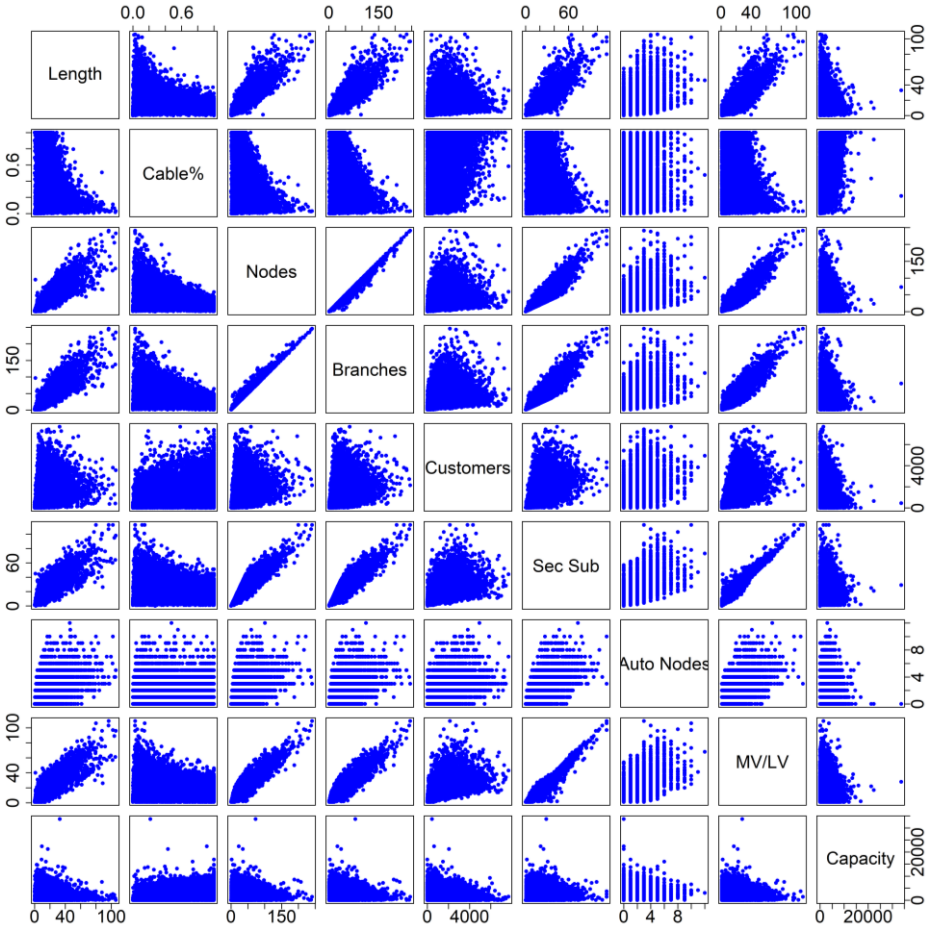


Figure 2.2 Scatter plot of numeric features of 15kV feeders

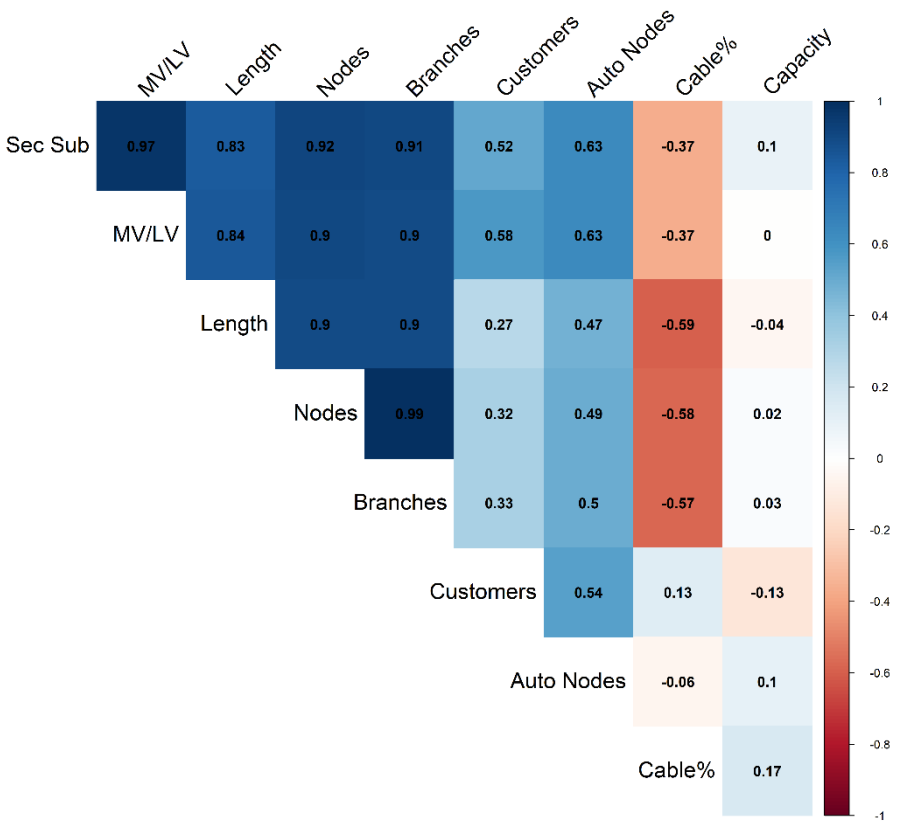


Figure 2.3 Pearson correlation coefficients between numeric features

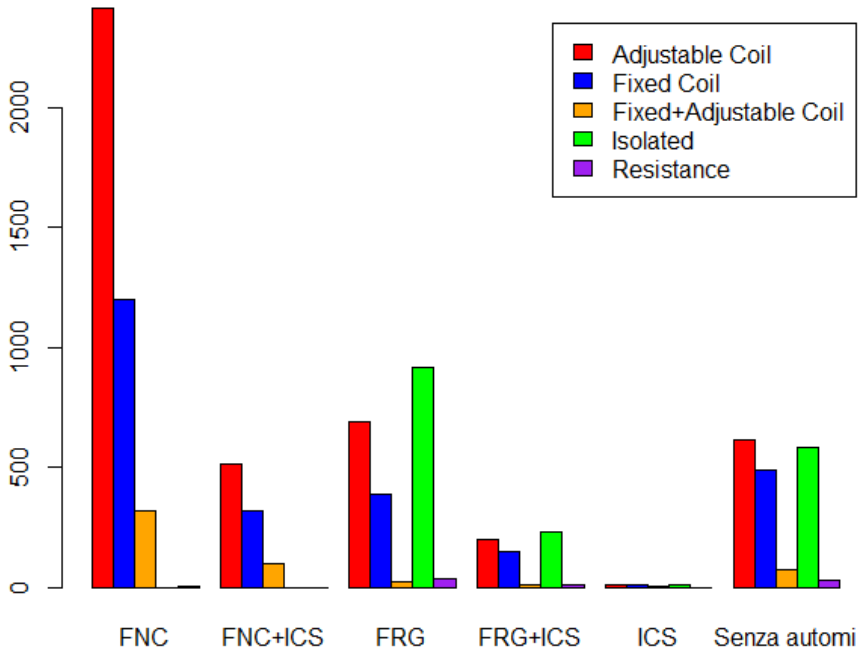
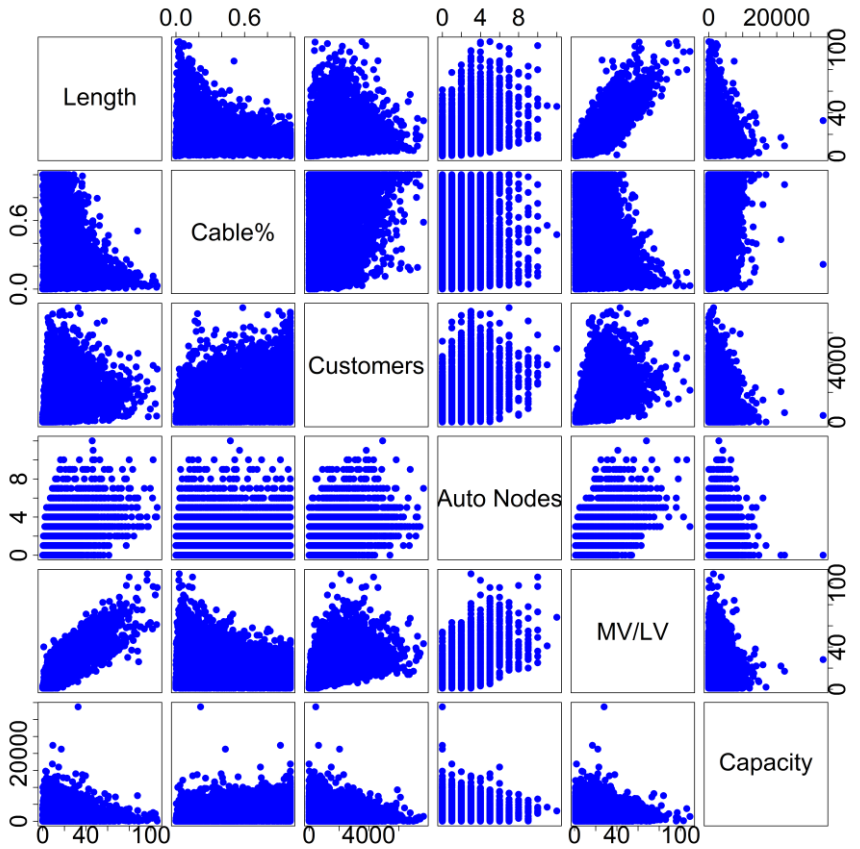
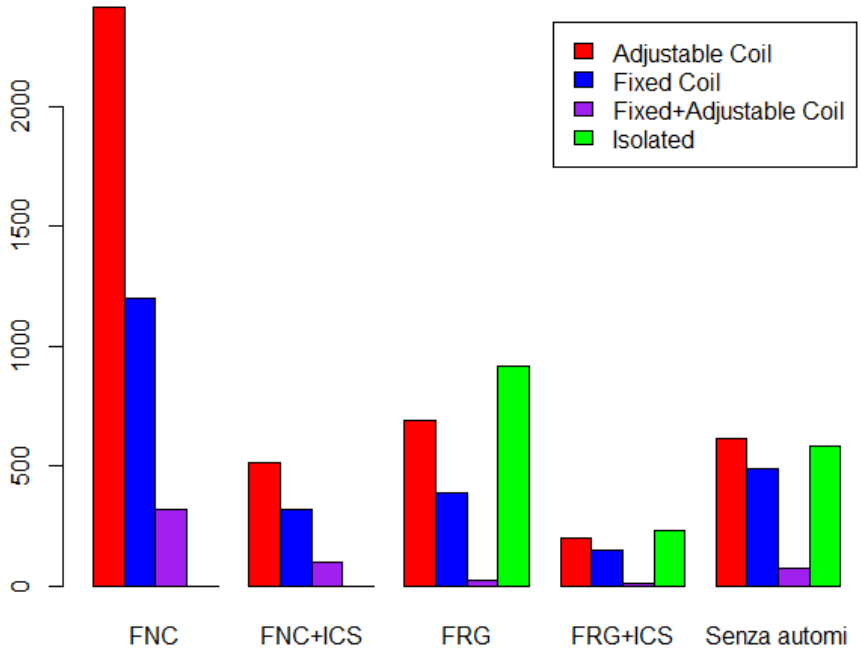


Figure 2.4 Neutral grounding types for each Automation Type

After eliminating the highly correlated numeric features and low frequency categorical features, there are 9243 MV feeders left with their features shown in Figure 2.5.



(a) Numerical features

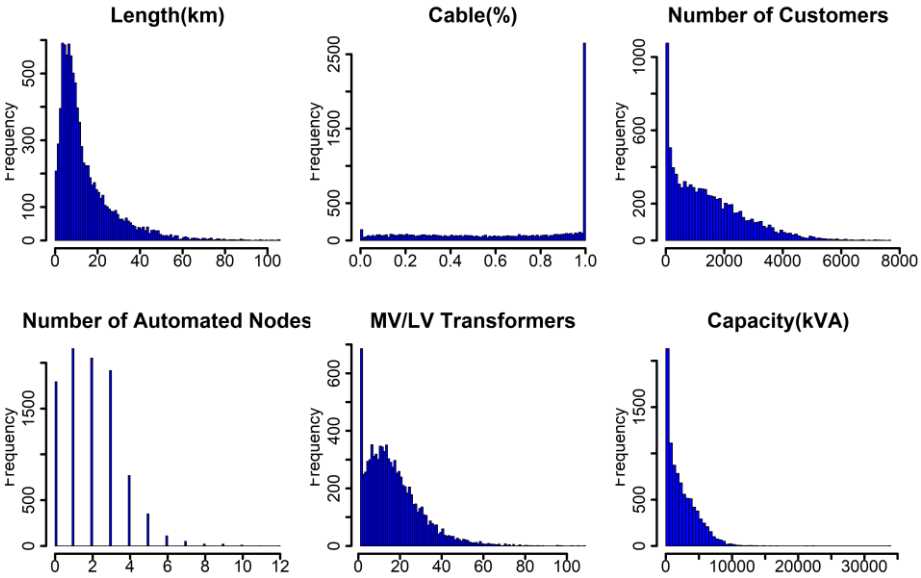


(b) Categorical features

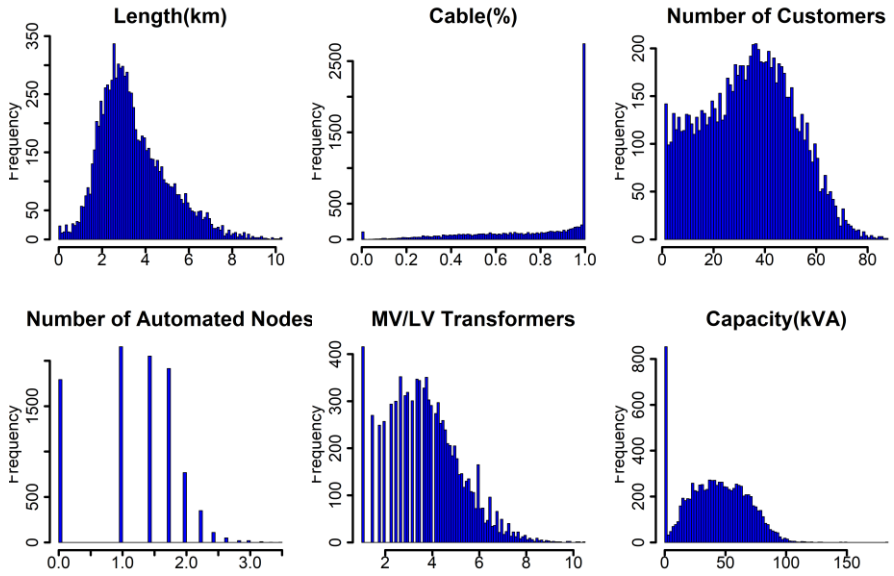
Figure 2.5 Selected features for clustering algorithm

2.2.3 Distribution of numeric features

As a method to partition data into groups based on the similarity between data points, the clustering technique is sensitive to the skewed data distributions, i.e., imbalanced data [138]. Figure 2.5(a) shows the distributions of selected numeric features among 9243 MV feeders. It is obvious that most of, much of distribution lines are about 10 km with less than 20 MV/LV transformers. Only few feeders are longer than 60 km and serve more than 4000 customers. Since these skewed-distributed data may lead the clustering method to poor results [139], the square root method is an effective way to modify the highly skewed distribution as shown in Figure 2.6(b). The number of outliers is dramatically dropped.



(a) Before square root



(b) After square root

Figure 2.6 Distribution of numeric features in MV feeders

2.3 Clustering process

Since a large number of MV feeders are spread in a wide spatial space with a diversity of structural features, it is difficult and time-consuming to evaluate the performance of each line. To get a better understanding of the thousands of MV feeders, clustering method is adopted as a powerful data mining technique to find out groups with similar structural feeders. This method breaks the dataset of total MV feeders into groups with minimum sum of differences between feeders labelled in the same cluster and maximum differences between clusters [140-141]. As an unsupervised learning algorithm, the feature values as well as centroid of each group are determined automatically. Both partitioning around medoids (PAM) and hierarchical clustering algorithms are to be used in the clustering analysis as the representative approaches.

Suppose there are n objects (X_1, X_2, … X_n) with p attributes in the data set X as shown in (2.2).

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1(p-1)} & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2(p-1)} & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n(p-1)} & x_{np} \end{bmatrix} \tag{2.2}$$

where x_{ip} is the p-th feature value of object X_i, p is the total number of features. There are n objects in the data set.

A dissimilarity measure between two objects (X_i and X_j) drawn from the same feature space must be chosen carefully for the clustering process. The Minkowski metric is a common method to evaluate the dissimilarity between objects as shown in (2.3). When m=2, the Minkowski metric is in fact the Euclidean distance, which is most popular for continuous numeric features.

$$d_m\left(X_i, X_j\right) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^m \right)^{\frac{1}{m}} = \left\| x_i - x_j \right\|_m \tag{2.3}$$

However, before doing the clustering algorithm, there are two special acts need to be taken in calculating the dissimilarity:

Although (2.3) can be used to calculate the distance between objects in numeric feature space, Figure 2.6 reveals that the value ranges of these features are not in the same scale, which may lead to a highly biased result. For example, the maximum of cable percentage is 1, while the number of customers is far larger than 1 in most cases. This nearly eliminates the significance of cable percentage in the clustering analysis. To make the selected features in a comparable range, the normalization approach is applied on the numeric features of selected data according to

$$x_{ip,norm} = \frac{x_{ip} - x_{p,\min}}{x_{p,\max} - x_{p,\min}} \tag{2.4}$$

where $x_{p,min}$ and $x_{p,max}$ is the minimum and maximum value of the p-th feature for all the objects, respectively. After normalization, the numeric feature values of objects are successfully constrained in the range [0,1].

For categorical values in a data set, the dissimilarity measure between two objects can be defined by the mismatches of the corresponding features [142-143], which is known as the Gower's distance as shown in (2.5)

$$d\left(X_i, X_j\right) = \sum_{k=1}^{p} \delta\left(x_{ik}, x_{jk}\right) \tag{2.5}$$

where

$$\delta\left(x_i, x_j\right) = \begin{cases} 0 & \left(x_i = x_j\right) \\ 1 & \left(x_i \neq x_j\right) \end{cases}$$

It is straightforward to integrate the two dissimilarity definitions together for a mixed data set as in (2.6)

$$d\left(X_i, X_j\right) = \sum_{k=1}^{c} \delta\left(x_{ik}, x_{jk}\right) + \sqrt{\sum_{k=c+1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{2.6}$$

where c is the number of categorical features and Euclidean distance is adopted in the numeric feature space.

2.3.1 Clustering algorithm

A. PAM algorithm

PAM is an effective clustering algorithm which has an advantage in identifying homogeneous groups of objects from large data sets. This algorithm aims at searching for k representative objects as medoids in the data set. Each cluster is constructed with one medoid and the nearest data points around it. The best k medoids will achieve the minimum sum of the dissimilarities of observations to their closest representative object.

- *Step1*: Initializing the medoids. To make an initial guess at the centers, the dissimilarity between each pair of objects need to be computed according to (2.6). The initial medoids can be determined by calculating (2.7) at each object and then sorting them in ascending order [144]. Q_j reflects the overall distance between object j and the other objects to some extent. The first k objects with minimum values are taken as the initial cluster medoids.

$$Q_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \tag{2.7}$$

- *Step2*: Assign the each of the rest objects to its nearest medoids and calculate the sum dissimilarity from objects to their medoids in each cluster.

- *Step3*: Replace the current medoid with new one in each cluster to minimize the sum dissimilarity in a cluster.
- *Step4*: Reassigning all the objects in data set with the new medoids and calculate the sum dissimilarity from objects to their medoids in each cluster. If the value is equal to the previous one, the optimal clustering results are already presented. Otherwise, go back to Step 3.

B. Hierarchical algorithm

Although Pam algorithm works well in most cases, it needs a specific number of clusters at the beginning. Hierarchical clustering is an alternative approach which aims to build a hierarchy of clusters because the clustering results are presented in a dendrogram. For an agglomerative hierarchical clustering algorithm, the starting number of clusters is the same as the number of objects and each cluster contains only one object. Then the pairs of clusters are merged as one moves up the hierarchy until all objects are merged together in a unique cluster.

- *Step1*: Regard all the data points as individual clusters and calculate the proximity matrix of these clusters.
- *Step2*: Merge the two closest clusters and update the proximity matrix.
- *Step3*: Repeat Step2 until all the clusters are merged together and form a single cluster.

2.3.2 Clustering evaluation

According to the algorithms introduced above, the clustering results can be derived with any number of clusters no larger than the total number of objects in data set. Therefore, the quality of clustering results needs to be evaluated with different cluster numbers.

A. Average silhouette coefficient

Silhouette coefficient is a technique to evaluate how well an object is assigned to this cluster. It is defined as (2.8) with a range of [-1,1].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.8}$$

For each object i, $a(i)$ is the average dissimilarity between i and the other objects in the same cluster; $b(i)$ is minimum average dissimilarity between object i and all the objects in any other cluster. A large value of $s(i)$ means that the object i is assigned to a much compact cluster while a negative value of $s(i)$ indicates that the object i may be assigned to a wrong cluster. The average silhouette coefficient of all the objects in a data set can be used to evaluate the quality of clustering results.

B. Calinski-Harabasz index

The Calinski-Harabasz index is defined as (2.9)-(2.11)

$$c = \frac{B/(k-1)}{W/(n-k)} \tag{2.9}$$

$$B = \sum_{i=1}^{k} n_i \left| G_i - G \right|^2 \tag{2.10}$$

$$W = \sum_{i=1}^{k} \sum_{j \in n_i} \left| X_j - G_i \right|^2 \tag{2.11}$$

where B is the sum of squared distances between cluster centroids G_i and global centroid G, multiplied by the number of objects n_i in the corresponding cluster. W is the sum of distances between all the objects in data set and their own cluster centroids. The number of total objects and clusters are denoted with n and k, respectively. Similar to the average silhouette coefficient, a larger value of Calinski-Harabasz index indicates a longer distance between clusters and shorter distance between objects in the same cluster.

2.3.3 Clustering Results

In our research, the 9234 MV feeders with 8 features are taken for clustering analysis. By applying PAM algorithm with cluster number from 2 to 40, the cluster index for each object can be obtained. While for the hierarchical clustering, the result is a dendrogram which can be cut into groups as shown in Figure 2.7. A dendrogram is a tree-like diagram that records the sequences of merges through the steps of agglomerative hierarchical clustering algorithm.

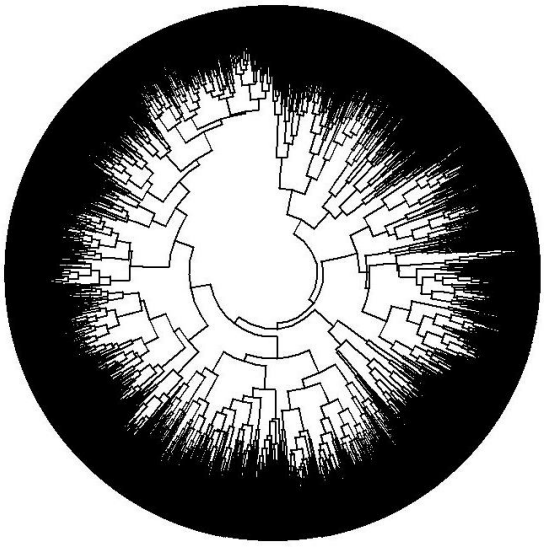
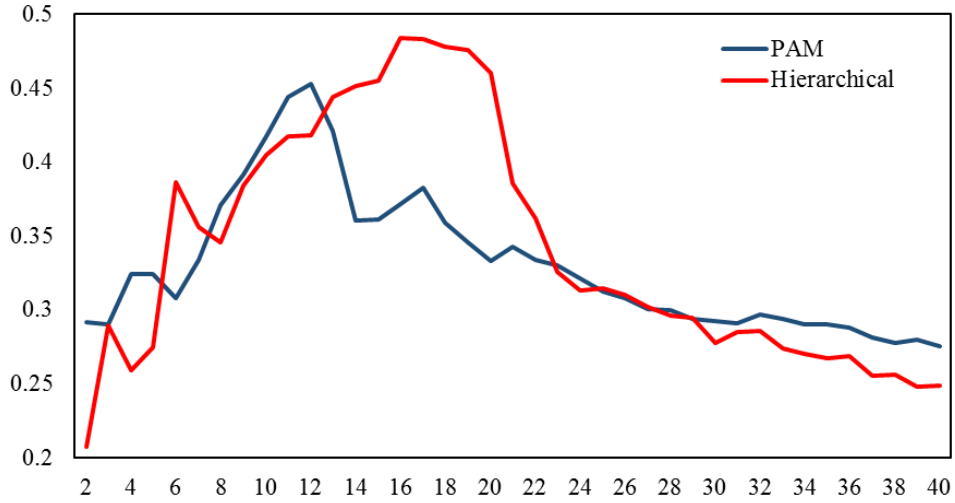


Figure 2.7 Hierarchical clustering of 15kV feeders

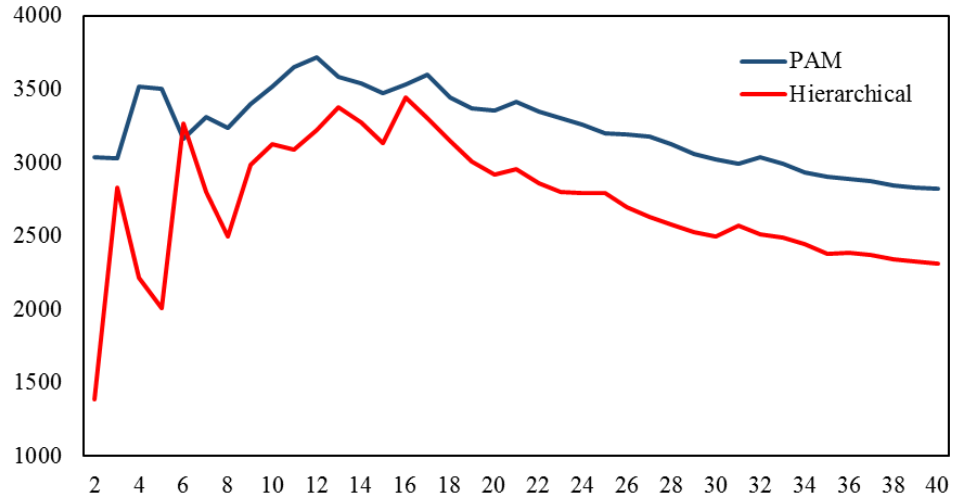
A. Best clustering number

As mentioned before, the average silhouette coefficient and Calinski-Harabasz indices are effective indicators to evaluate the clustering result. Both indexes can be regarded as the quotient of distance between groups divided by the

compactness of inside group objects. Figure 2.8 demonstrates the trend of these indexes with increasing number of clusters.



(a) Average silhouette coefficient for different number of clusters



(b) Calinski-Harabasz index for different number of clusters

Figure 2.8 Evaluation indices for clustering results

As shown in Figure 2.8, the average silhouette coefficient and Calinski-Harabasz index have a good consistency for the best cluster numbers. For PAM algorithm, the highest values of both indexes appear when cluster number is 12. The same situation happens for hierarchical clustering only when cluster number is 16, which can be plotted on the dendrogram with each color for a cluster in Figure 2.9. Although the best average silhouette coefficient for PAM algorithm is lower than that for hierarchical clustering method, the Calinski-Harabasz index shows in a contrary way. It is difficult to determine which algorithm is better for clustering results.

B. Visualization of clustering results

Another efficient way to evaluate the quality of clustering results is to map the high-dimensional data into a low-dimensional space. The t-distributed stochastic

neighbor embedding technique is a powerful method for nonlinear dimension reduction [145-146]. It can model each MV feeder in the data set by a two-dimensional point, which is then visualized in a scatter plot.

The t-distributed stochastic neighbor embedding method is an effective way of nonlinear dimensionality reduction in the topic of manifold learning. The principle of stochastic neighbor embedding is to first construct a probability distribution between high-dimensional data that allows a higher probability of being selected between the close data objects; then, a low-dimensional space (2-D for visualization) is constructed by affine transformation, whose probability distribution is as close as possible to the former one.

As for the Hierarchical clustering, the similar objects are supposed to have a closer distance similar to those shown in Figure 2.9.

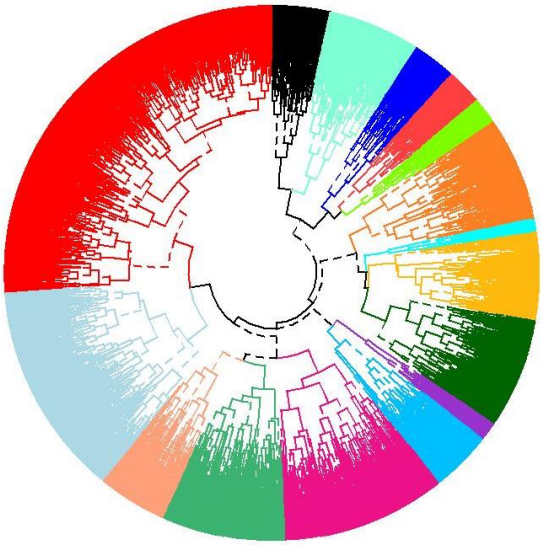
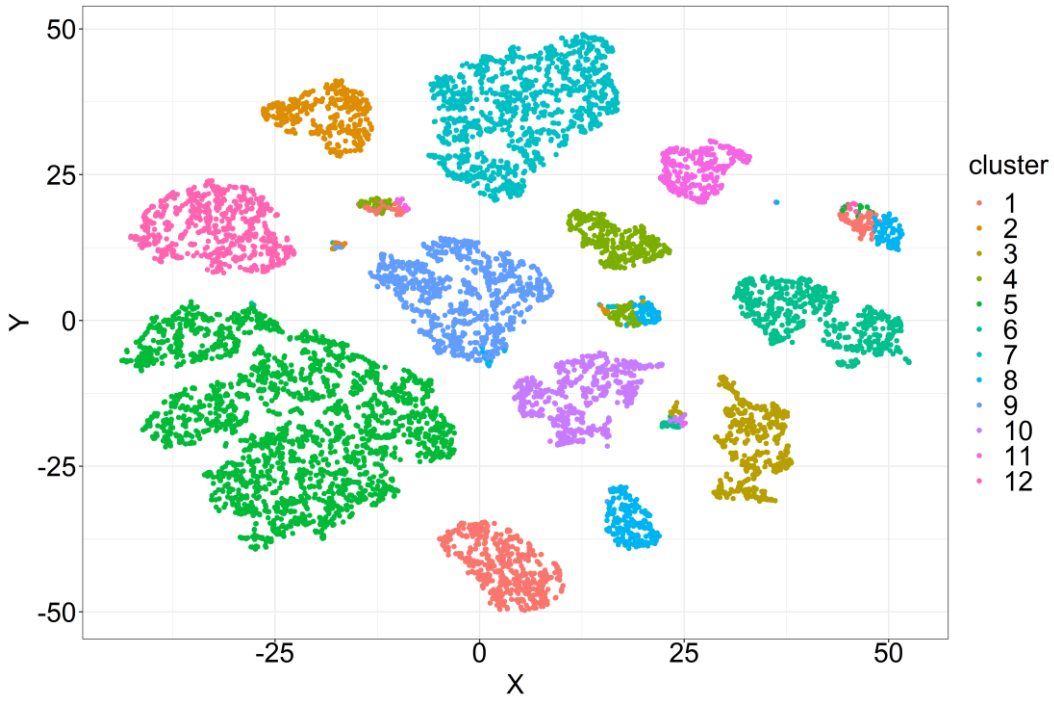
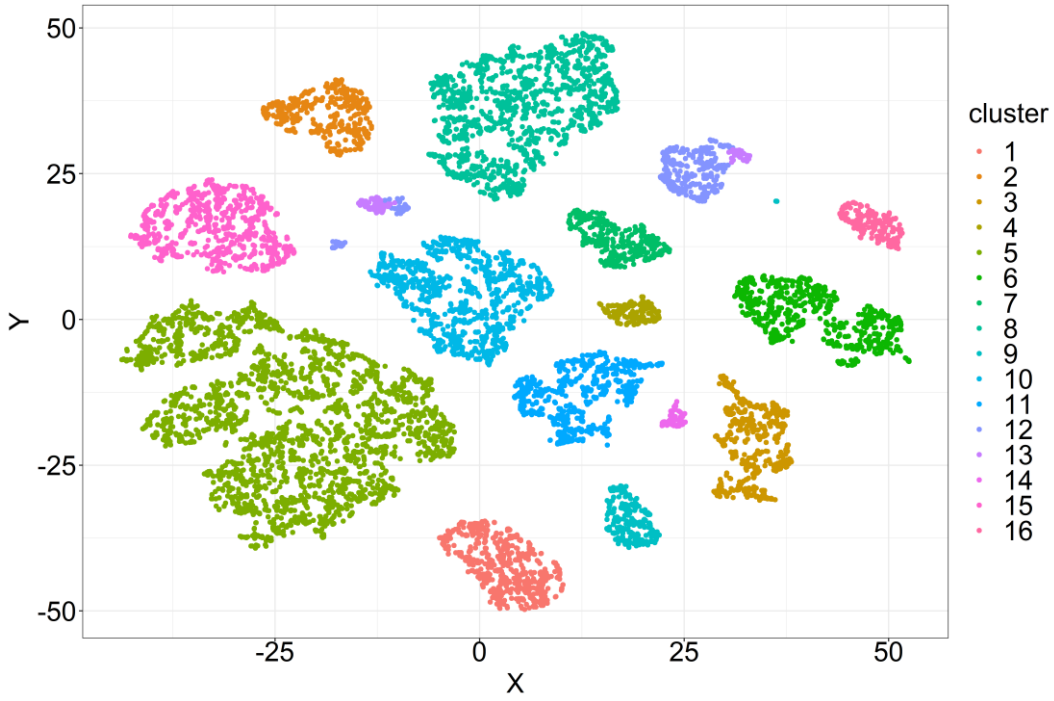


Figure 2.9 Dendrogram of 16 clusters with hierarchical clustering

In Figure 2.10, the 9243 MV feeders are mapped into a two-dimensional scatter plot with each color indicating an independent cluster. As is shown in Figure 2.9, most of clusters have been well recognized by the two methods. Since the best cluster number of hierarchical clustering is larger than that of the PAM algorithm, there are some small-scale groups can also be identified without much confusion. But as a kind of dimension reduction technique, the scatter plot in Figure 2.10 can only be taken as a supplementary means for results verification.



(a) PAM algorithm



(b) Hierarchical clustering

Figure 2.10 Visualization of different clusters in a low-dimensional space

C. Structural characteristics of each cluster

According to the best clustering results under PAM and hierarchical clustering method, the structural features of each cluster are displayed in Figure 10-13. The boxplots in Figure 2.11 and Figure 2.12 demonstrate the distribution of numeric features of each cluster under PAM and hierarchical clustering

methods, respectively. Since there is a strong correlation between MV/LV transformers and the secondary substations, nodes, as well as branches, the distribution of three latter variables are supposed to have the similar pattern with that of the transformers. As to the 4 neutral grounding modes and 5 automation types shown in Figure 2.13 and Figure 2.14, there exists only one dominant value in each cluster. Significant differences among clusters verified the adequate grouping results. Based on the median value of numeric features and dominant value of categorical features, the representative feeders of 9243 MV feeders under the above two methods can be derived as in Table 2-3 and Table 2-4.

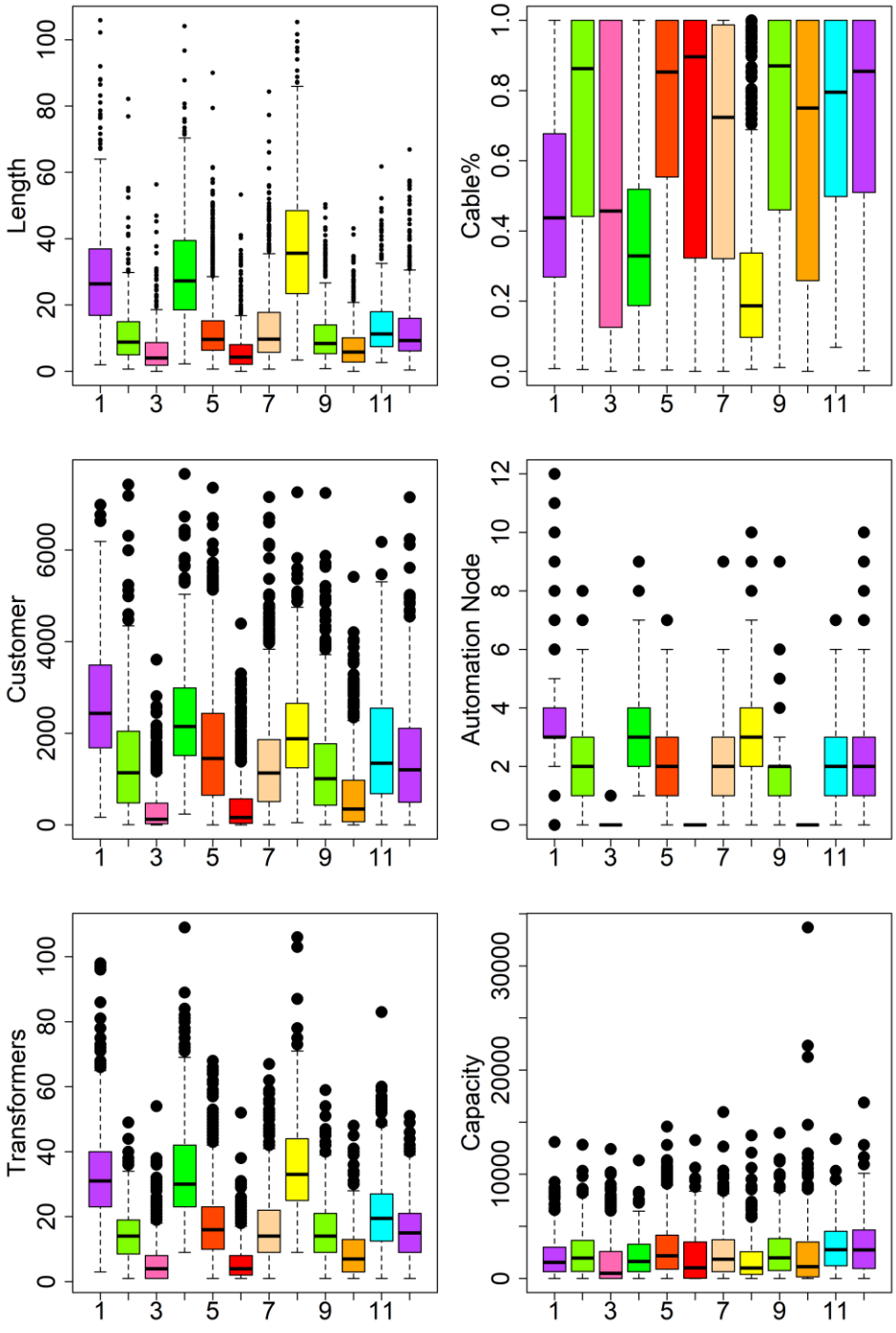


Figure 2.11 Numeric feature distribution with PAM algorithm

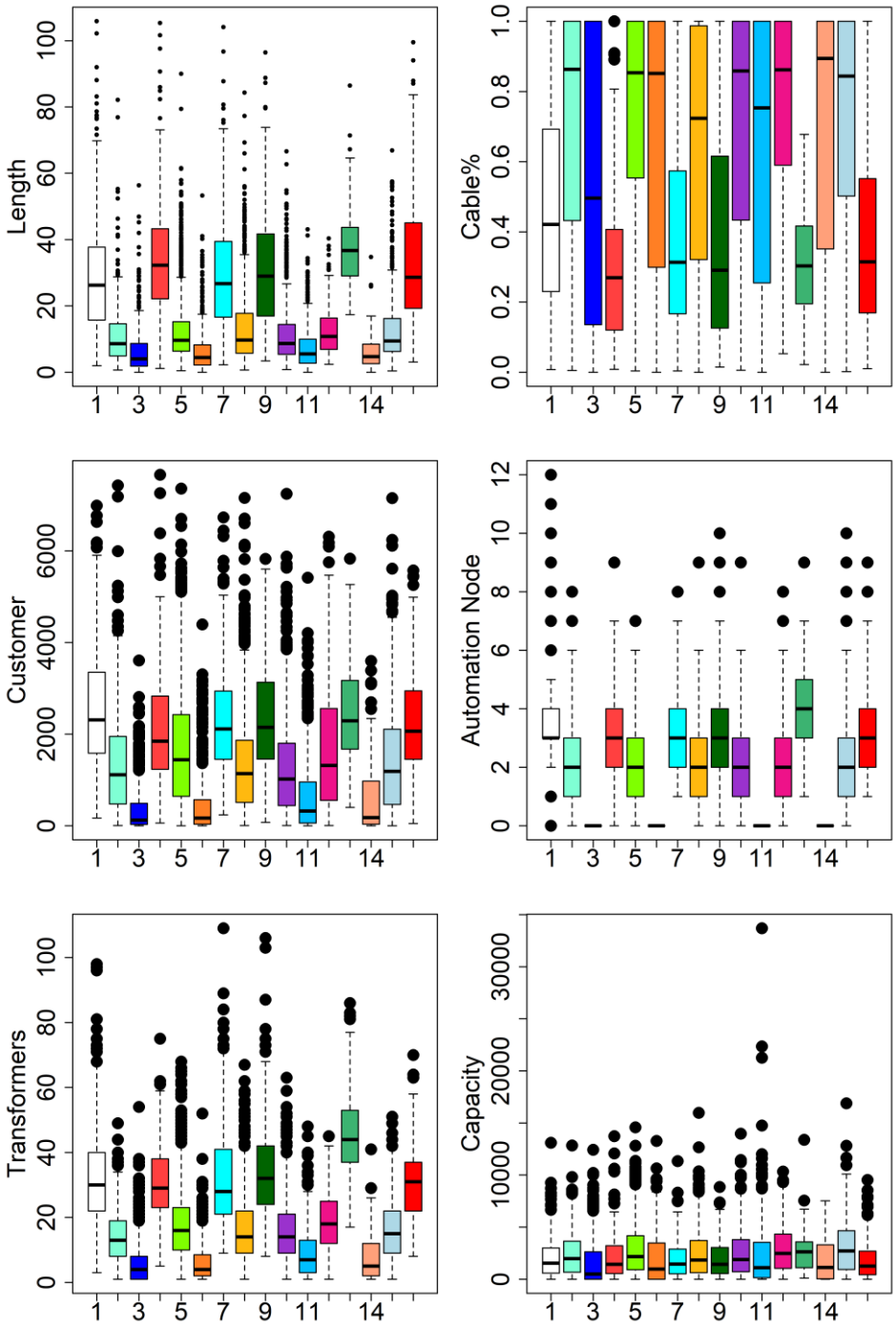
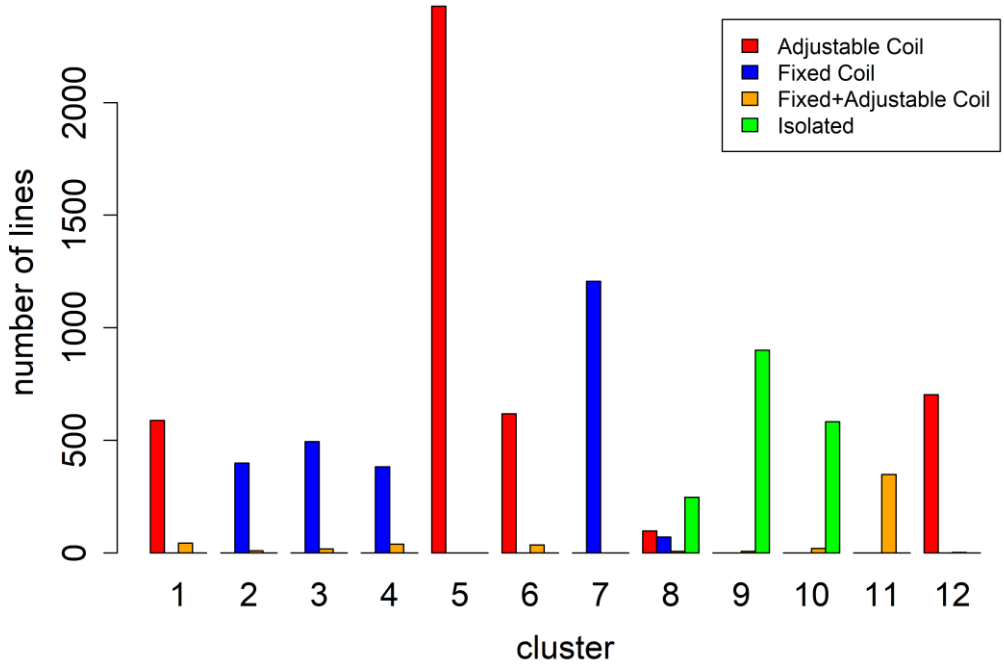
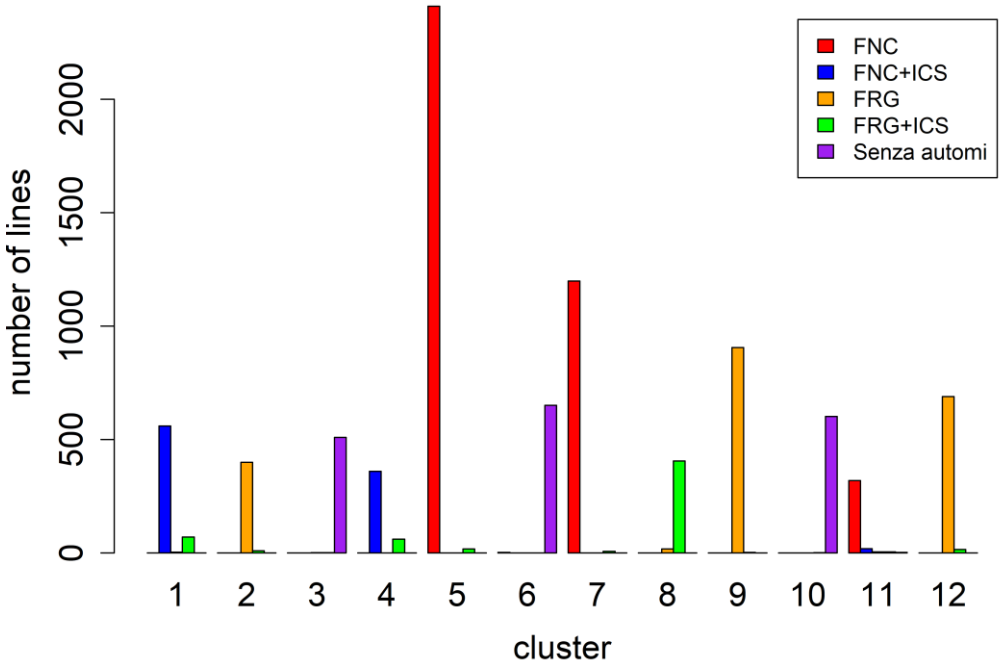


Figure 2.12 Numeric feature distribution with Hierarchical algorithm

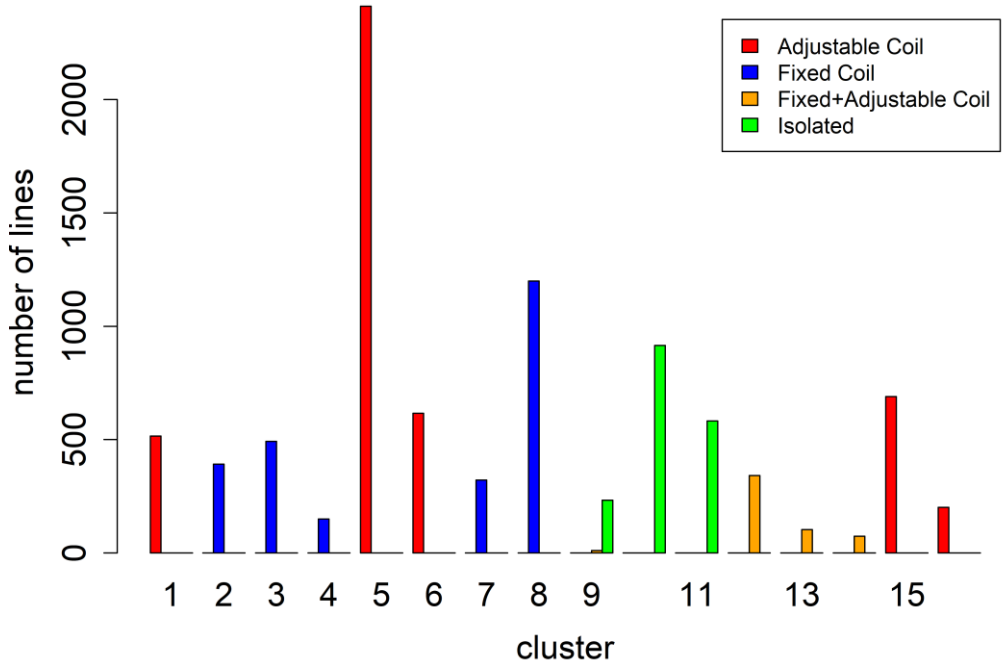


(a) Neutral grounding mode

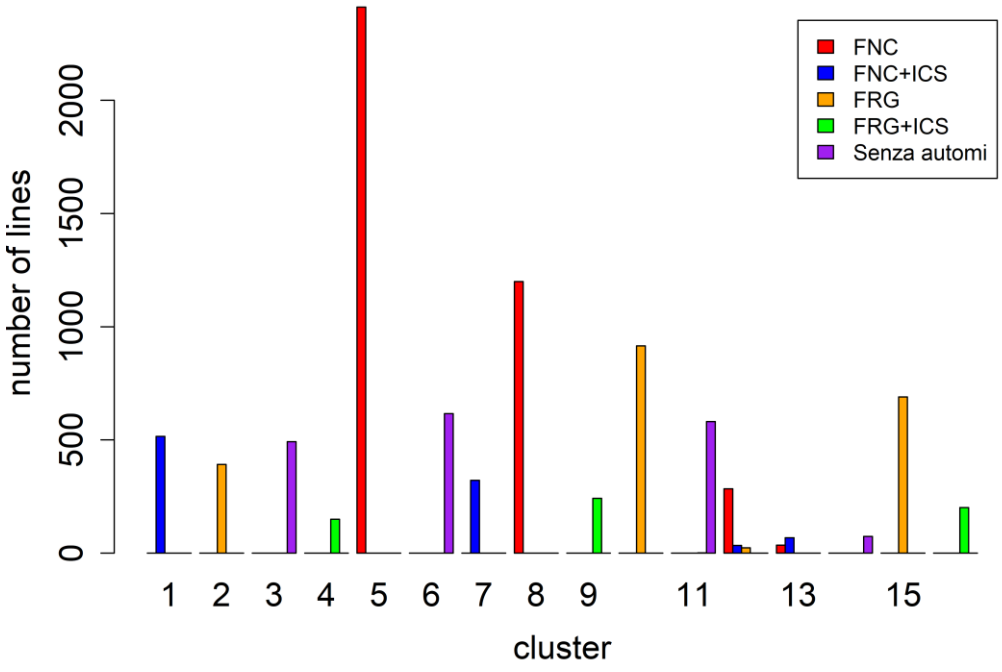


(b) Automation Types

Figure 2.13 Categorical feature distribution with PAM algorithm



(a) Neutral Grounding Mode



(b) Automation type

Figure 2.14 Categorical feature distribution with Hierarchical algorithm

Table 2-3 Representative Feeders under PAM Algorithm

	Length (km)	Cable (%)	Nodes	Branches	Customers	Secondary Substation	Auto Nodes	Transformer	Capacity (kVA)	Neutral Mode	Automation
Clu1	26	44	55	61	2438	34	3	31	1541	Adjustable Coil	FNC+ICS
Clu2	9	86	20	24	1137	16	2	14	1651	Fixed Coil	FRG
Clu3	4	46	8	9	122	5	0	4	487	Fixed Coil	None
Clu4	27	33	64	71	2149	35	3	30	1625	Fixed Coil	FNC+ICS
Clu5	10	85	22	25	1449	19	2	16	2159	Adjustable Coil	FNC
Clu6	4	89	7	8	163	5	0	4	1025	Adjustable Coil	None
Clu7	10	72	23	25	1136	17	2	14	1644	Fixed Coil	FNC
Clu8	36	19	70	75	1883	37	3	33	985	Isolated	FRG+ICS
Clu9	8	87	20	22	1010	16	2	14	1983	Isolated	FRG
Clu10	6	75	12	14	347	8	0	7	1119	Isolated	None
Clu11	11	80	28	31	1347	22	2	20	2758	Fixed+Adustable Coil	FNC
Clu12	9	85	22	25	1202	17	2	15	2732	Adjustable Coil	FRG

Table 2-4 Representative Feeders under Hierarchical Clustering

	Length (km)	Cable (%)	Nodes	Branches	Customers	Sec Substation	AutoNodes	Transformer	Capacity (kVA)	Neutral Mode	Automation
Clu1	26	42	54	60	2310	34	3	30	1542	Adjustable Coil	FNC +ICS
Clu2	9	86	20	23	1116	15	2	13	1976	Fixed Coil	FRG
Clu3	4	50	8	9	122	5	0	4	504	Fixed Coil	None
Clu4	32	27	65	73	1845	33	3	29	1422	Fixed Coil	FRG +ICS
Clu5	10	85	22	25	1443	19	2	16	2170	Adjustable Coil	FNC
Clu6	5	85	7	9	167	5	0	4	968	Adjustable Coil	None
Clu7	27	31	60	64	2114	33	3	28	1460	Fixed Coil	FNC 10+ICS
Clu8	10	72	23	25	1138	17	2	14	1839	Fixed Coil	FNC
Clu9	29	29	61	68	2149	37	3	32	1439	Isolated	FRG +ICS
Clu10	9	86	20	23	1018	16	2	14	1897	Isolated	FRG
Clu11	5	75	12	13	323	8	0	7	1103	Isolated	None
Clu12	11	86	25	29	1318	21	2	18	2474	Fixed+Adjustable Coil	FNC
Clu13	37	30	85	93	2293	49	4	44	2623	Fixed+AdjustableCoil	FNC +ICS
Clu14	5	89	8	10	178	6	0	5	1121	Fixed+Adjstable Coil	None
Clu15	9	84	22	25	1187	17	2	15	2722	Adjustable Coil	FRG
Clu16	29	31	59	66	2066	32	3	31	1261	Adjustable Coil	FRG+ICS

2.4 Performance of each cluster

Based on the interruption records of all the 15kV feeders in 2014, the percentage of interrupted lines in each cluster can be calculated and displayed in Figure 2.15 and 2.16. Total number of feeders in each cluster is also added at the abscissa axis. Compared with different clusters in each figure, a lower percentage of interrupted lines indicates a better performance for one cluster. The result from PAM algorithm shows that cluster 3 and cluster 6 have the best performance with failure rate lower than 40%. As can be seen from Figure 2.11, the common characteristics between cluster 3 and 6 includes the short length of feeders as well as lower number of customers and MV/LV transformers. Since the length of representative feeders is only 4 km, there is no automation equipment adopted. The neutral grounding mode of cluster3 and 6 also indicates that the adjustable and fixed coils have little difference in reducing the number of interrupted lines. But it may have an impact on the number of repeated failures for each feeder. The similar situation also happens in the cluster 3, 6 and 14 under hierarchical clustering method.

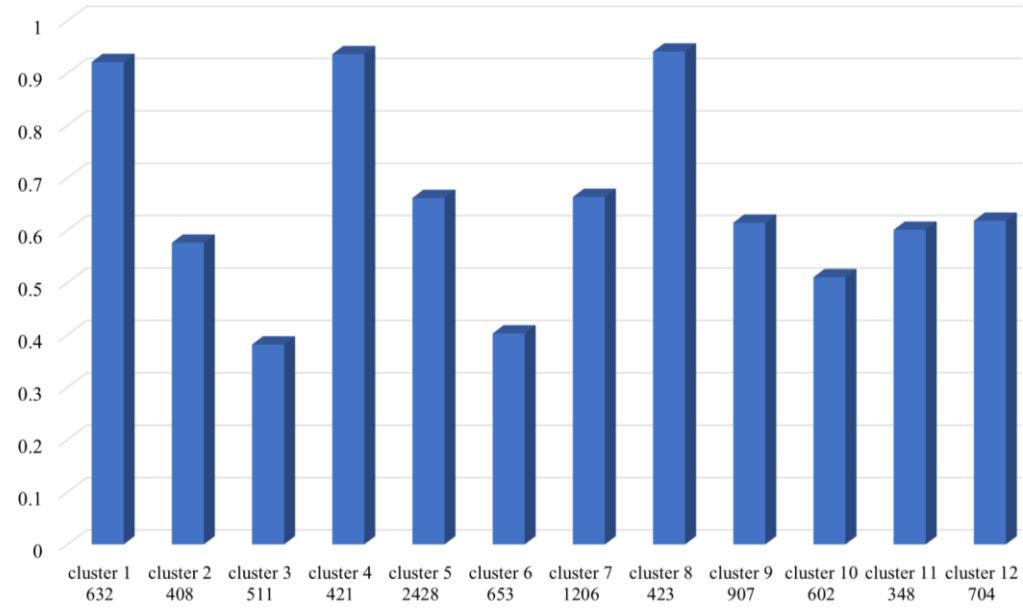


Figure 2.15 Percentage of interrupted feeders in each cluster (PAM)

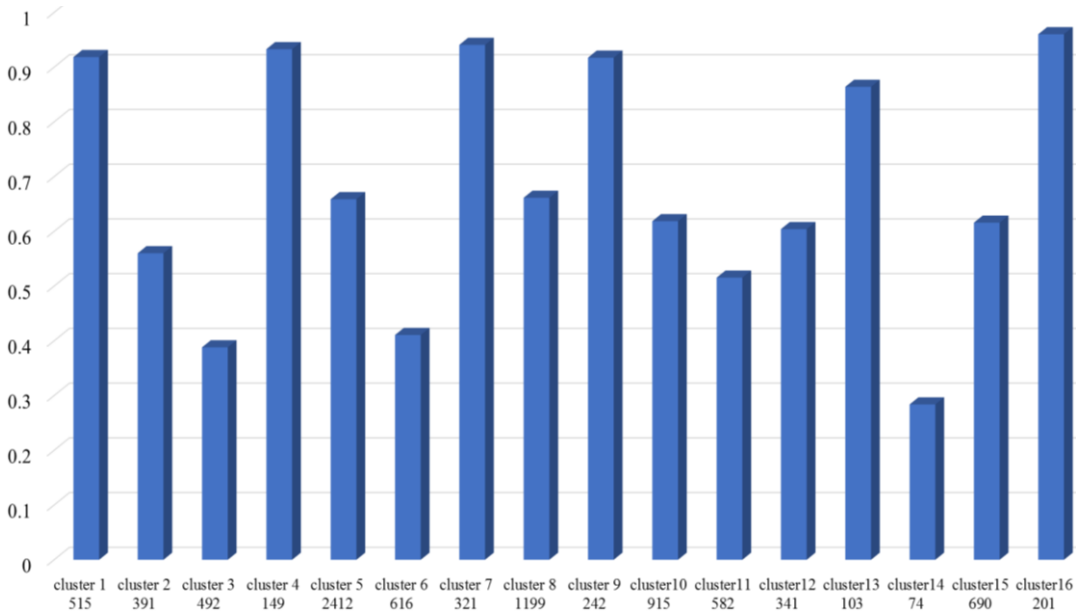


Figure 2.16 Percentage of interrupted feeders in each cluster (Hierarchical clustering)

Figure 2.15 also shows that the most common feeders are classified into cluster 5, with the medium value of numeric features as well as adjustable coil in the neutral grounding point and FNC automation. Most of these feeders are also successfully classified into cluster 5 under hierarchical clustering method, which verified the consistency between the two methods.

2.5 Summary

To classify tens of thousands of 15 kV feeders into typical patterns, this work applied the clustering method on the feeder structural features. Since some of the features are not independent, they are firstly selected based on the correlation analysis. To cope with the mixed data, Gower's distance is calculated to measure the dissimilarity between two feeders. PAM algorithm and hierarchical clustering method are both adopted for the analysis.

According to the clustering results, the dominant categorical features and value range of numeric features in each cluster are derived. The diversity between clusters and uniformity of each cluster indicates the effectiveness of both clustering methods. Although the best number of clustering results are different, there is a high consistency between these two methods. The representative feeders are selected as the medoids of each cluster.

Based on the interruption data of all feeders, the rate of interrupted lines in each cluster are calculated, which can be taken as a performance evaluation for each representative feeder. Compared with different feeders and their performance, the effect of automation type and impact of customers or length can be analyzed, which is essential for the distribution network planning.

Chapter 3

Prediction of outages in distribution network

3.1 Grey theory and outages number prediction

As the terminal of power system, the distribution network is characterized with the multi-type lines, wide area and complex structure. Due to this nature, the occurrence of interruptions has various causes. In general, there are three main factors for distribution system outages: intrinsic factors, like the defects of equipment; external factors, like the damages from animals or lightning; and human error factors [147]. However, utilities usually focus more on the repair of power system components rather than detailed investigation of failure causes. Therefore, the outage records contain only limited information for tracing back and analyzing the causes.

As one of the densest distribution networks, the Turin area suffers an increasing number of outages on the underground medium voltage grid. The forecasting of number of outages is an important reference for the DSO to anticipate the maintenance staff and evaluate the repairing cost from an operational point of view [148]. With more frequent heat waves, aging equipment and growing power demand, the outage number is increasing under multiple reasons. The grey model prediction method could capture the increasing trend of outages avoiding detailed analysis of factors affecting the performance of distribution network [149].

Outage management is critical for distribution system operators (DSO) as a long-standing problem [150]. There are various researches aiming at damage assessment of the power grid with statistical analysis [151-152]. These methods usually combine the historical power outage records with the environmental data as a forecasting model. However, in the urban distribution grid, most of the medium voltage feeders are underground cables, which brings the difficulties to analyze the relation between the external factors and the failures of underground

power feeders. The ability to forecast the number of outages in a period is important for the DSO to arrange the maintenance plans and improve the grid resilience [148, 153]. The annual outage prediction based on the records in a few years is also challenging due to the limited data and uncertainty.

The grey theory, which is proposed to solve the uncertain problems with rare or inadequate data [154], has seen various applications to the engineering problems [155-156]. Different from the typical forecasting models like Autoregressive Integrated Moving Average (ARIMA), the grey theory could overcome the limitation of insufficient data collections for calibrating the model parameters. Moreover, the ability to deal with sparse data enables it as a practical and user-friendly forecasting method. A grey model is utilized in [157] for predicting the failure rate of power substation in service to improve the safety and reliability of the equipment. To deal with the uncertainty of failure rate prediction in both fault stabilization period and fault loss period, a grey-linear regression model is used in [149]. The seasonal grey model, PSO-based grey model and the adaptive parameter learning mechanism based seasonal fluctuation grey model are built and compared in [155] to analyze the seasonal fluctuations of the electricity consumption of the primary economic sectors. A rolling mechanism designed on the principle of "new information priority" is combined to an optimized grey prediction model to forecast the electricity consumption in [156].

Among the forecasting models based on the grey theory, the GM(1,1) model is the most popular one with one variable and its one-order equations. The annual number of outages in urban distribution network can be predicted with the grey theory model because there is inadequate information about the factors in records for outages.

In typical grey models, the background values are defined as the average of two accumulative values. In this paper, the particle swarm optimization (PSO) method and genetic algorithm (GA) will be introduced to determine the optimized weights for calculating the background value with an objective as the minimum error between simulation results and the real values. Compared with the typical models, the PSO-based grey model and GA-based grey model improve the accuracy of the prediction and achieves satisfactory performance.

3.1.1 Grey prediction model

A. Basic grey model

Grey model is a classic method for studying the trend from discrete data series with limited samples and inadequate information. By accumulating the original data series, the randomness between samples could be reduced and a clear trend is possible to be revealed. Therefore, the first step of grey model is to add up the values.

In grey model, the data are denoted as vector $\mathbb{X}^{(0)}$ with a superscript (0) for original dataset. The original data vector contains N data points from $\mathbb{X}^{(0)}(1)$ to

$\mathbb{X}^{(0)}(N)$. The inherent properties and the trend among the original data points can be uncovered with the grey prediction model, whose procedures are as follows:

- Create a new vector $\mathbb{X}^{(0)}$ by accumulating the first k elements of original data $\mathbb{X}^{(0)}$

$$\mathbb{X}^{(1)}(k) = \sum_{i=1}^{k} \mathbb{X}^{(0)}(i) \quad k=1, \ldots, N \tag{3.1}$$

  where the superscript (1) refers to the accumulated data.

- Build the background value as Z(k)

  The k-th background value $Z(k)$ is defined as the average value of the k-th and (k-1)-th accumulated data as shown in (3.1).

$$\mathbb{Z}(k) = 0.5\mathbb{X}^{(1)}(k-1) + 0.5\mathbb{X}^{(1)}(k) \tag{3.2}$$

  where k=2, …, N.

- Estimate the model parameters a and b

  The first-order grey differential equational of a single variable is given in eq. (3) [156]

$$\mathbb{X}^{(0)}(k) + a\mathbb{Z}(k) = b \tag{3.3}$$

  where k=2, …, N.
  The corresponding white differential equation is as follows:

$$\frac{\mathrm{d}\mathbb{X}^{(1)}(t)}{\mathrm{dt}} + a\mathbb{Z}(t) = b \tag{3.4}$$

  where a and b are the development coefficient of the system and endogenous control grey scale, respectively.
  By rewriting the above equation in the matrix format, we could get the following

$$\begin{bmatrix} \mathbb{X}^{(0)}(2) \\ \mathbb{X}^{(0)}(3) \\ \vdots \\ \mathbb{X}^{(0)}(n) \end{bmatrix} = \begin{bmatrix} -\mathbb{Z}(2) & 1 \\ -\mathbb{Z}(3) & 1 \\ \vdots & \vdots \\ -\mathbb{Z}(n) & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} \tag{3.5}$$

  With the known original data $\mathbb{X}^{(0)}$ and background values $\mathbb{Z}$ available, the parameters a and b could be estimated by the least square method.

- Derive the estimation values of the accumulated data

  Once the parameters a and b of the grey prediction model are obtained, the direct output of the model is the estimation of accumulated values as in (3.6)

$$\widehat{\mathbb{X}}^{(1)}(k+1) = \left(\mathbb{X}^{(1)}(1) - \frac{b}{a}\right) e^{-ak} + \frac{b}{a} \tag{3.6}$$

The vector $\hat{\mathbb{X}}^{(1)}$ is the estimation of vector $\mathbb{X}^{(1)}$ with (n-1) values.

- The prediction of grey theory model

The estimation of the original data is determined with (3.7)

$$\hat{\mathbb{X}}^{(0)}(k+1) = \hat{\mathbb{X}}^{(1)}(k+1) - \hat{\mathbb{X}}^{(1)}(k) \tag{3.7}$$

B. PSO-based GM(1,1)

One of the effective measures to improve the accuracy of typical grey model is to calculate the background value $\mathbb{Z}(k)$ with optimized weights on two continuous values as in the following equation:

$$\mathbb{Z}(k) = \mu\mathbb{X}^{(1)}(k-1) + (1-\mu)\mathbb{X}^{(1)}(k) \tag{3.8}$$

where μ is 0.5 in the typical GM(1,1) indicating an averaged number of the two cumulative values $\mathbb{X}^{(1)}(k)$ and $\mathbb{X}^{(1)}(k-1)$. However, with an optimized weight μ, the proposed model is possible to capture more details of the trend hidden in the original data. Since it is difficult to specify a clear formula for calculating the value of μ, finding the optimal value of the weight could not be accomplished by the traditional optimization methods. Instead, the heuristic optimization algorithms, like the PSO algorithm, will yield a pretty good result.

- PSO algorithm

The PSO algorithm is initialized with a group of random solutions and seeks for the optima in the given domain. In the algorithm, each solution to the optimization problem is regarded as a bird, or a "particle" in a more general way. The procedures of PSO algorithm is inspired from the stimulation of the foraging behavior of birds. Therefore, with the number of dimensions as D, there are two vectors to indicate the position X_i and velocity V_i of the i-th bird, respectively.

$$X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$$
$$V_i = (v_{i1}, v_{i2}, \dots, v_{iD})$$

In order to find the optimal solution, each "particle" is evaluated in the objective function with a fitness value given their positions. Then, the velocity of each particle is to be adjusted according to its own and the others' searching experience. By substituting the position of the i-th particle into the objective function, the fitness of this particle at the current iteration is calculated. If the optimization problem is to find the minimal solution for the objective function, the position corresponding to the minimum fitness value among all the k iterations of the i-th particle is regarded as the personal best solution $pbest_i^k$. Similarly, the position corroding to the minimum fitness value of all the particles among the k iterations is denoted as $gbest^k$ for the global optimal solution. For a D-dimensional optimization problem, both $pbest_i^k$ and $gbest^k$ are vectors containing D elements.

The position of each particle cannot be directly changed in the PSO algorithm. Instead, it is adjusted with the velocity for the next iteration, whose value is updated with the following equation:

$$v_{id}^{k} = wv_{id}^{k-1} + c_1 r_1 \left(pbest_{id} - x_{id}^{k-1}\right) + c_2 r_2 \left(gbest_d - x_{id}^{k-1}\right) \quad (3.9)$$

where d indicate the d-th dimension of the velocity (d=1, 2, …, D), w is the inertia factor; c_1 and c_2 are the personal and group learning factor, respectively; r_1 and r_2 are the random numbers in the range [0,1].

Then, the d-th dimension of the position of the particle i is updated with the following formula:

$$x_{id}^{k} = x_{id}^{k-1} + \alpha v_{id}^{k} \quad (3.10)$$

where α represents the weight for the velocity.

The terminal condition of the program is usually set as the maximum number of iterations K or minimum criteria of the errors.

- PSO-based GM(1,1)

In the grey model GM(1,1), the weight μ for calculating the background values $\mathbb{Z}$ could be optimized with the PSO algorithm. The objective of the optimization is to minimize the errors between the predictions and real values. In this research, the Mean Absolute Percentage Error (MAPE) is used to evaluate the performance of the grey prediction model, which is defined as below:

$$\text{MAPE} = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{\mathbb{X}^{(0)}(k) - \hat{\mathbb{X}}^{(0)}(k)}{\mathbb{X}^{(0)}(k)}\right| \quad (3.11)$$

The procedures for the PSO-based grey model prediction are summarized in Figure 3.1.

C. GA-based GM(1,1)

- GA algorithm

The GA optimization [158] originated from the theory of natural evolution and mimics the process of natural selection, including mutation, crossover and selection. It starts with a population of randomly distributed solutions in the given domain. Each candidate solution corresponding to the value of μ differs from the others. In the GA, the candidate solutions are called chromosomes represented by n-bit "genes".

There are two 14-bit chromosomes composed by binary genes shown in Figure 3.2. In our case, the value of μ is within the range [0,1], therefore the real values of the candidate solutions can be decoded by locating the positions represented by the chromosomes in [0,1]. The difference of the real values between the two nearest chromosomes refers to the precision of the final solution with GA. With 14-bit chromosomes, the minimum real value of the difference between two candidate solutions is $6.11*10^{-5}$, which indicates that the precision of the solution is 4 decimals.

In our case, the objective of the optimization process is to minimize the errors between the estimated values and real values. Therefore, the Mean Absolute Percentage Error (MAPE) is used as the objective function. Then the performance of each solution could be evaluated by the reciprocal of objective function as the fitness.

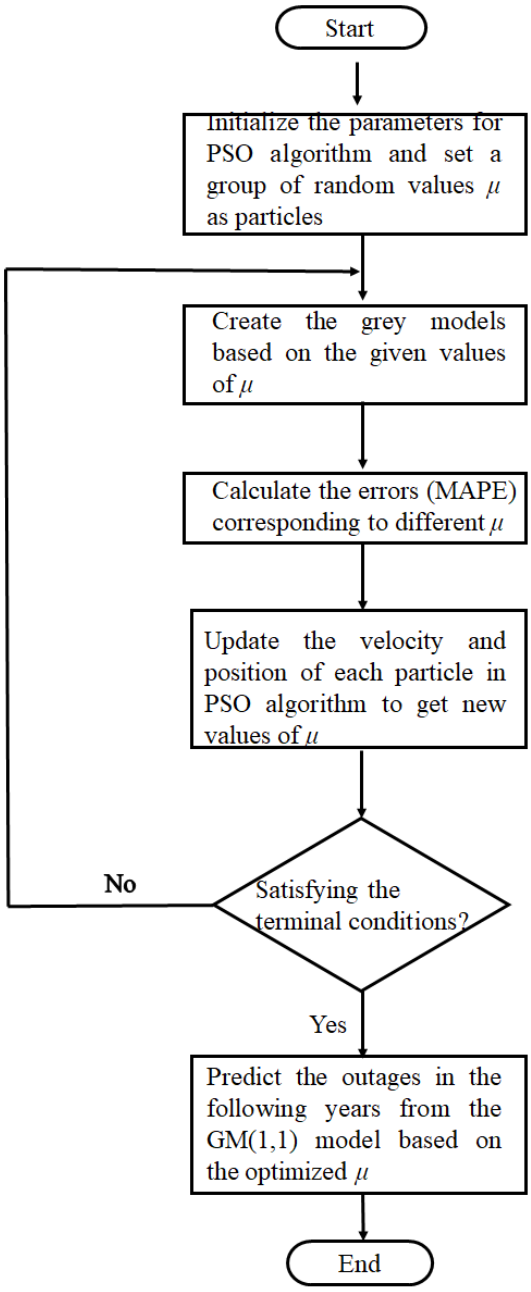


Figure 3.1 Procedures of PSO-based GM(1,1)

In the selection step, these chromosomes are picked up with a probability proportional to their fitness values. Therefore, candidate solutions with a less value of objective function (i.e., with higher fitness) are more likely to be selected for the following steps.

As for the crossover procedure, the selected chromosomes from the previous step are further combined by exchanging part of their binary strings to each other.

In order to create more variability, a smaller rate is typically set for the mutation step.

After all the above steps, the fitness of each individual will be calculated again in the new iteration until the stopping criterion is reached. If the variation of the objective function from the two successive iterations is less than a pre-defined small threshold, the algorithm is considered as converged. Otherwise, the iteration process will be terminated until it reaches the maximum number. The schematic diagram of the GA-based grey model is shown as Figure 3.3.

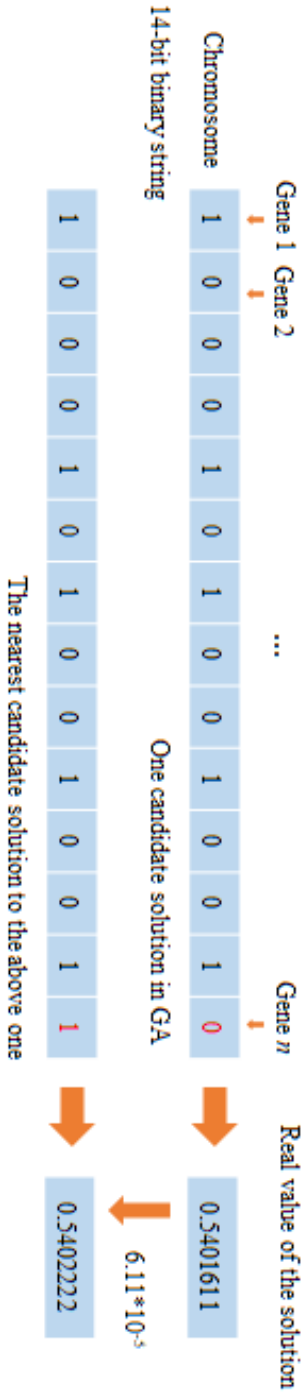


Figure 3.2 Procedures of PSO-based GM(1,1)

- Rolling mechanism for GA-based GM(1,1)

In practical cases, the trend inside a data sequence is typically significant in a limited period. Since only the latest data reflect the development trend, the rolling mechanism with a sliding window for the building of GM(1,1) is an effective method to deal with the seasonal data.

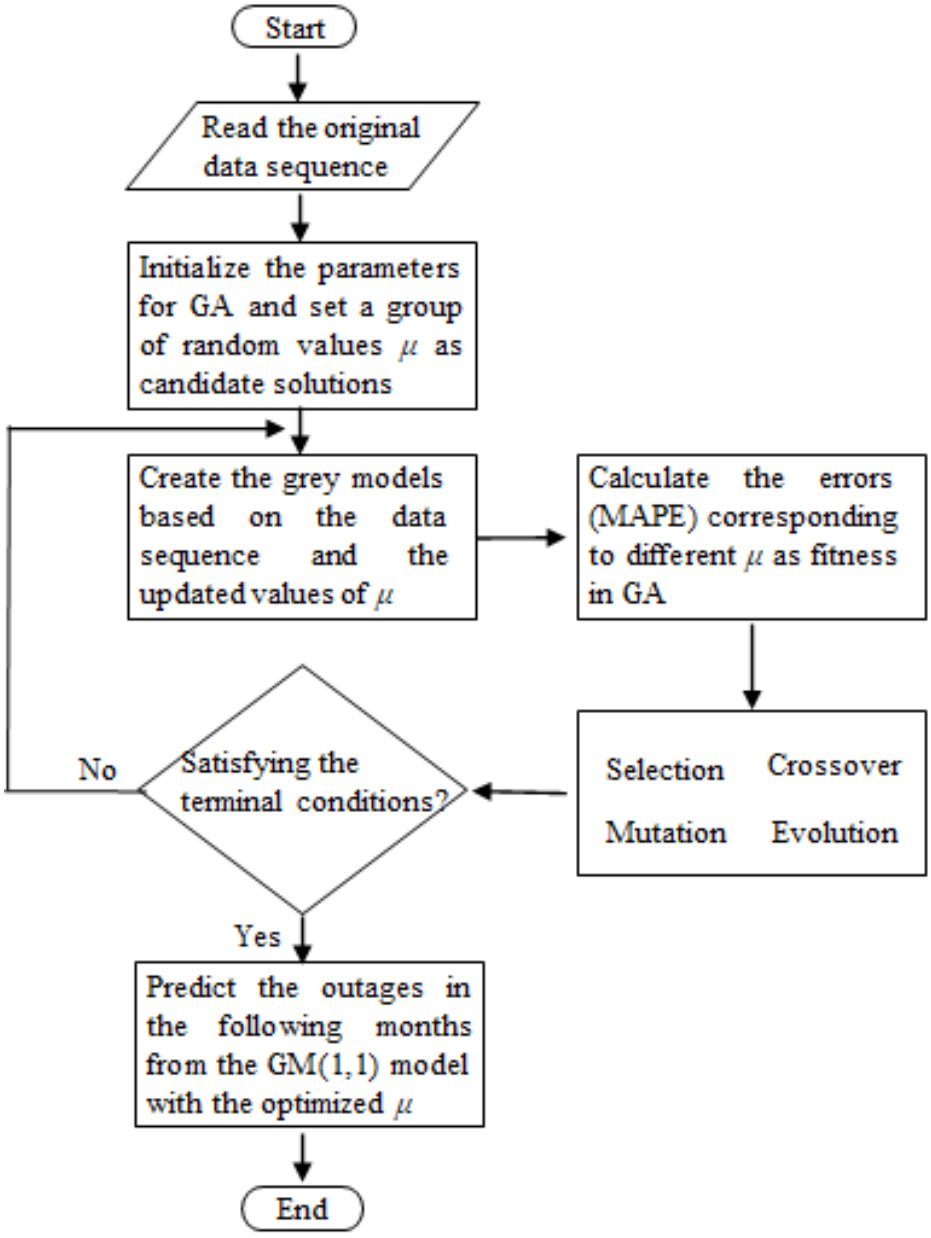


Figure 3.3 Procedures of GA-based GM(1,1)

In the rolling mechanism, the GA-based GM(1,1) is always built on the p original data points $\left[x_1^{(0)}\ x_2^{(0)}\ \ldots\ x_p^{(0)}\right]$ to predict the following q data points $\left[x_{p+1}^{(0)}\ x_{p+2}^{(0)}\ \ldots\ x_{p+q}^{(0)}\right]$. In this case the length of sliding window is p. Once the new information is acquired, the predicted q data points will be replaced with the real values. Then most updated p real values $\left[x_{q+1}^{(0)}\ x_{q+2}^{(0)}\ \ldots\ x_{p+q}^{(0)}\right]$ will be utilized to predicted the next q values $\left[x_{p+q+1}^{(0)}\ x_{p+q+2}^{(0)}\ \ldots\ x_{p+2q}^{(0)}\right]$. These steps will be literately implemented over all the data set.

3.1.2 Case Study

A. Annual outage number prediction

In practice, the distribution network is planned in accordance with the development of the city. The capacities of the substations and feeders are usually updated with the changes of the layout of residential and industrial areas. Meanwhile, the aging process of power equipment also bring difficulties to the outage analysis in a long-time framework. Therefore, the outage number recorded within 10 years is taken as a proper time span for our research.

According to the local utility, the outage records of the urban distribution system in Turin from 2008 to 2014 is shown in Fig. 3.4. In this figure, the outage number of distribution network from 2008 to 2014 is shown as blue bars. The red curve connecting the top of each bar indicates the trend of outage numbers among the 7 years. As can be seen from this figure, in spite of some fluctuations, there is an increasing trend of the outages. In this work, a grey prediction model is built with these 7-year records according to the procedures in Fig. 3.1.

In the first step of PSO-based GM(1,1), the parameters in the PSO algorithm are supposed to be initialized. In this work, the population of particles is set as 20 and the maximum number of iterations are both set as 50, which means that if the algorithm cannot get converged in 20 iterations, it would be automatically terminated. The value of personal learning factor c1 is set as 0.25. In order to emphasize the group's impact and avoid the local optima, the group learning factor c2 is set as twice the value of c1. The maximum velocity is set as 0.1, which is 10% of the total range of μ.

Then, 20 random values distributed in [0,1] are fed to the PSO algorithm as the initial values of μ. For each iteration, the errors between the output of GM(1,1) and the real outage numbers are calculated, based on which both the personal best position of every particle and the group's best position could be determined. This information indicates the value and direction for the modification of next step's velocity. A new group of μ would be generated based on the particles' current position and new value of velocity. In our case, the PSO algorithm is converged within 20 iterations as shown in Fig. 3.5.

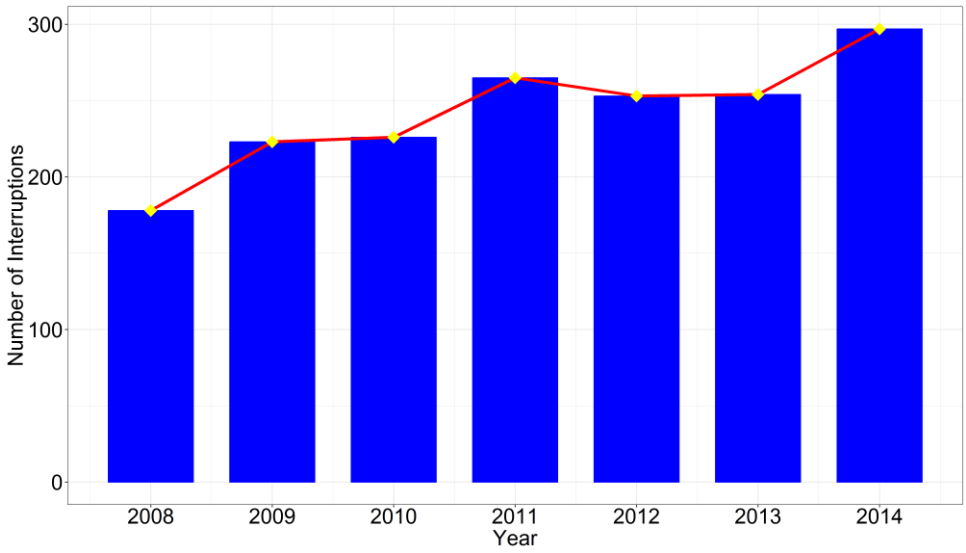


Figure 3.4 Number of outages in 7 years

In Figure 3.5, the MAPE between the prediction results from PSO-based grey model and the real values are demonstrated in each iteration. As a univariate optimization problem, the optimization process becomes converged very soon after a few iterations. The stable value of MAPE is 3.42%, implying a pretty good performance of the method. By applying the optimized μ into the grey model GM(1,1), the number of outages in in 2015 and 2016 can be predicted with the model derived above and compared with the results from GM(1,1) when μ=0.5, as listed in Table 3-1.

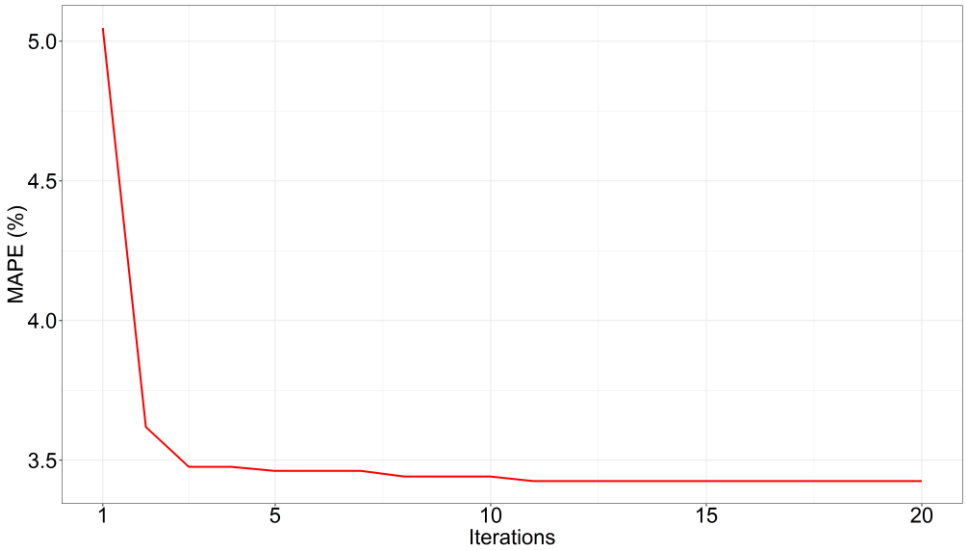


Figure 3.5 Value of MPAE with increasing number of iterations

Table 3-1 Prediction results from the grey model

		Optimized μ		μ=0.5	
Year	Real	Prediction	Error	Prediction	Error

	Number				
2015	296	301.6	1.89%	301.4	1.82%
2016	322	317.2	1.49%	315.7	1.95%

All the real outage numbers and the predicted values with PSO-based GM(1,1) from 2008 to 2016 are shown in Figure 3.6. One point needs to be emphasized is that, the parameter μ is optimized based on the previous years' records, which means if the outages in the next year changes dramatically, the optimized parameter may lead to a prediction a little far from the classic model (μ is always set as 0.5). This concept is also shown in Fig. 3.6, where the performance of optimized grey model is better than the non-optimized grey model in most cases, while it may behave a little worse in some other cases.

B. Monthly outage number prediction

From 2008 to 2012, the number of outages in every month is shown in Figure 3.7. As can be seen from the figure, the monthly outage number increases during summer and decrease in winter. The possible reason is that during summer, the higher temperature and less precipitation bring chanllenges to the insulation of power equipment. The aging process of electrical devices accelerates during such critical weather conditions. However, there is also a potential increasing trend in the annual total number of outages. If we focus only on the summer period, the outage number has seen a slight increase during the 5 years. It may be caused by the more frequent heat waves in summer. Meanwhile, the warmer summer encouges more families to install the air-conditioners in their houses, which significantly increase the load in power grid. Outages are prone to happen in the overload state.

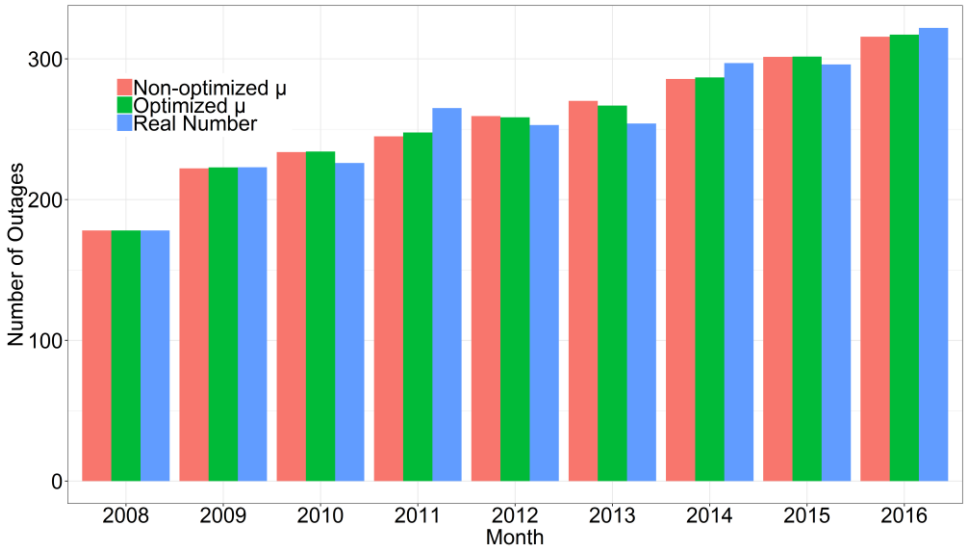


Figure 3.6 The number of annual outages from 2008 to 2016

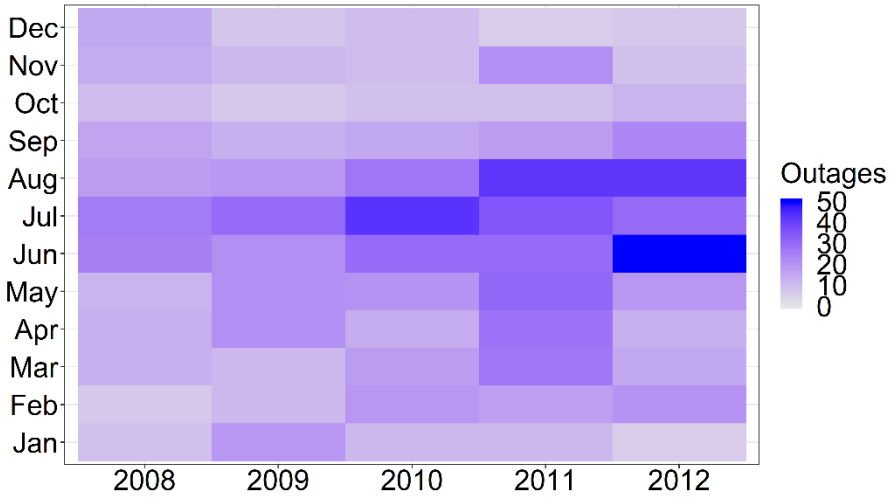


Figure 3.7 The number of Monthly outages from 2008 to 2012

Before using the rolling mechanism to evaluate the whole data set, we will build the grey models and test the feasibility of the proposed method with the first 7 months' records. Both the basic GM and GA-based GM are built based on the outage records from January to June. The number of outages in July is predicted. As can be seen from Figure 3.8, both the models successfully captured the trend among the first half year's records while the optimized μ in GA-based GM proved the improvement of prediction accuracy.

The monthly outage number in the urban distribution network among the five years will be predicted by empoying the GA-based GM(1,1) in this section. In order to capture the lateest trend of the change in outage numbers, the rolling mechanism is applied. The length of sliding window p will be tested on different values with comparison of forecasting errors. The length of prediction data q is always set as 1 in our simulation. The prediction results are shown in Figure 3.9.

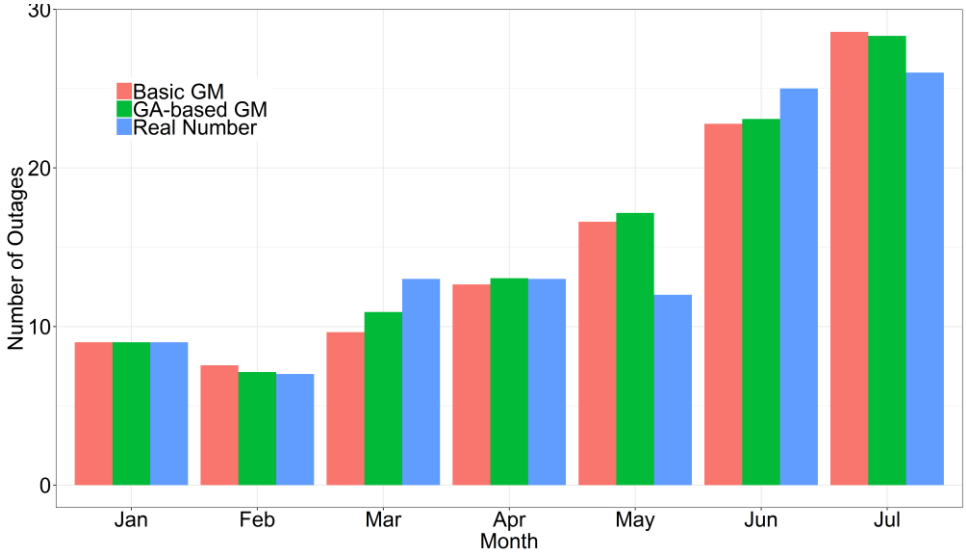


Figure 3.8 Results of GA-based grey model based on 7 months' records

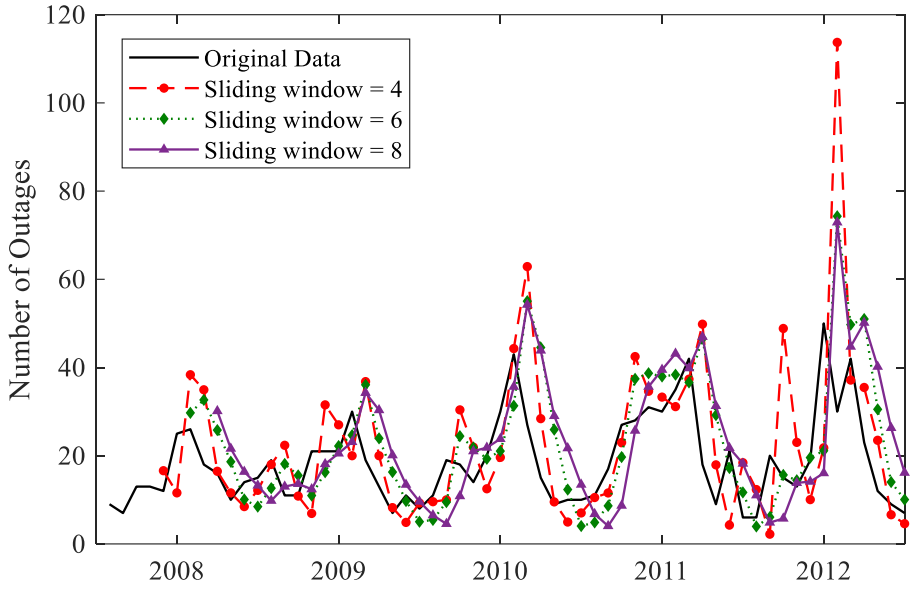


Figure 3.9 Results of monthly outage prediction with different sliding windows

3.2 SVM and Daily outages prediction

As the goal of utilities' pursuit, an uninterrupted power supply plays an important role in the customers' satisfaction and cost of operation. Therefore, a prior assessment of the outages becomes a practical problem for the distribution system operators in predictive maintenance and investment evaluation. In Europe, the underground cables are largely applied in the urban distribution network, which brings difficulties for analyzing the impact from the external factors.

In order to properly address the uncertainty of outages in distribution network, some researches have been carried out to reveal the causes of outages. The authors in [159] have proposed a method to estimate the outage rate of overhead transmission lines under wind storms. In their work, a fragility curve is developed to describe the relationship between the outage rate and the severity of wind storms. However, the wind storms are usually critical to the overhead lines in transmission system and has limited impact on the underground cables in distribution network. The authors in [160] built the weather condition dependent failure rate models based on the high-resolution radar observations of storm characteristics with a Bayesian outage prediction algorithm. As to the outages in distribution network, an ensemble learning approach is introduced in [161] to estimate the weather-caused power outages, especially those caused by the wind and lightning. The relation between duration of unplanned outages in distribution network and environmental factors is analyzed in [162] by learning from historical outage records.

In this section, our study will focus on the evaluation of daily outages in distribution network considering different weather conditions. The number of outages per day could be divided into two different critical levels: Level I with 0 or only 1 outage in the whole day, while the others are labeled as Level II. Apparently, Level II is more critical to the distribution system operators than

Level I. Those days identified as Level II need more attention for predictive maintenance and preparation of a real outage.

The potential impact of weather conditions on these two levels is first tested with the two-dimensional visualization from PCA. As an efficient dimension reduction method, PCA is good at converting the linear dependent dimensions in high dimensional data set into principal components. With more than half of information contained in the first two dimensions, the instances labeled as Level I and Level II could be visualized in a two-dimensional plot. The weather's impact on these two different levels could be roughly analyzed in an intuitive evaluation.

This task is then considered as a binary classification problem with different weather conditions as input and two levels of outage as output. Since most of the days are labeled as Level I in the real-world records, there is a serious imbalance in the data. In order to address this problem, an over-sampling method [163] is adopted to re-balancing the two levels in the data set. This method implements the over-sampling of the minority class by creating "synthetic" examples rather than by sampling with replacement. The synthetic examples are generated based on the original data according to specific rules, which could improve the classifier's performance to some extent. With the re-balanced data set, a classification model is to be built with the SVM algorithm. The ROC curve will be used to verify the validity of the proposed framework for evaluating the weather's severity

3.2.1 Data description

A. Dataset for analysis

Among the outage records between 2012 and 2017, the number of days without any outage and the total number of outages in every month are summarized and shown in Figure 3.10.

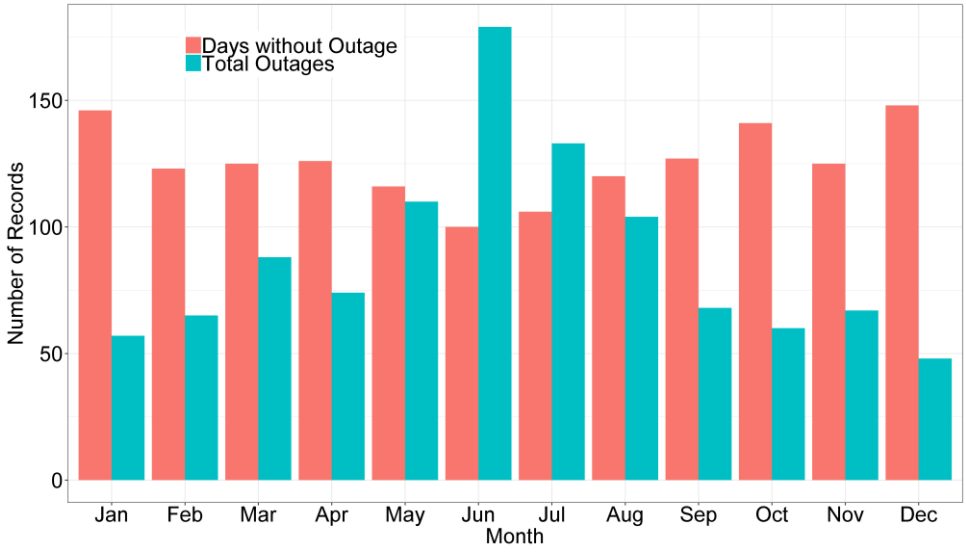


Figure 3.10 Outage records in different months

In the figure above, the value of each month is the sum of the records in the same month during the 6 years. Accordingly, there are about 180 days of each

month. As can be seen from Figure 3.10, all these months have at least 100 days without any outages, which indicates a reliable and secure power grid. However, a smaller number of days without outages are seen in summer, especially in June and July compared to the winter months. Similarly, the total number of outages in each month reaches the largest value in the summer time.

As for the local weather information, several original data are collected including the daily temperature, relative humidity, precipitation, and so on. The average value of daily maximum temperature in each month is shown in Figure 3.11.

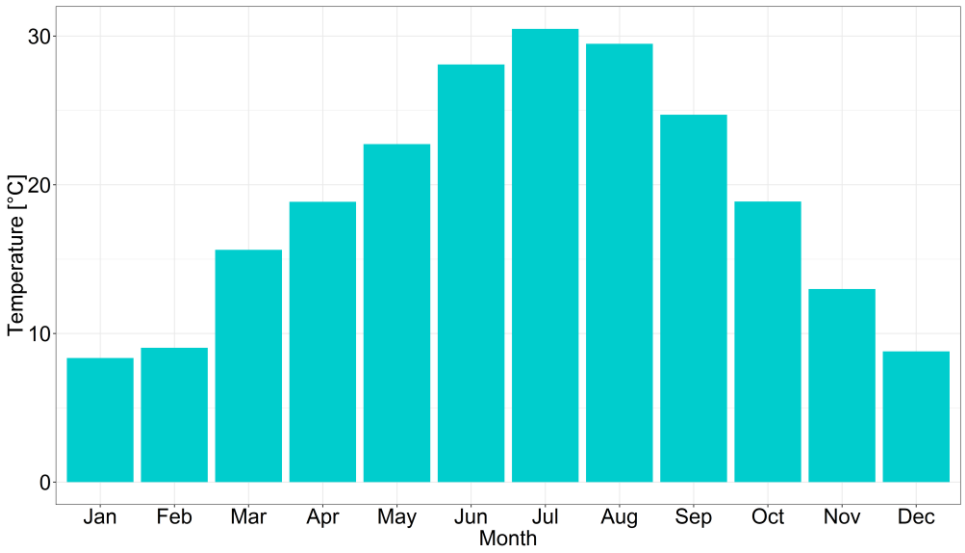


Figure 3.11 Average value of daily maximum temperature in different months

As is shown in Figure 3.11, the number of total outages in each month is approximately in cline to the variation of daily maximum temperature in different months. Moreover, the number of days with no outages decreases when the temperature is high in summer.

All these phenomena show a potential relation between the number of outages and the weather conditions. In order to investigate the weather's impact in detail, not only the daily extreme value, but also the average value in a period is taken into consideration for the continuous impact. In this study, apart from the three original daily weather features: maximum temperature, minimum relative humidity, total solar radiation and precipitation, we further calculated their average values in 5, 10, 15 days as 12 new features for analysis. The daily precipitation and cumulative value of precipitation in 30 days is also constructed as an important feature.

B. Data visualization

In this study, the severity of outages in distribution system is defined as two levels according to the number of outages per day. Among the outage records in 6 years, the days with more than two independent outages are labeled as "Level II",

which indicates a severe situation for special attention, and the rest of days are labeled as “Level I” as shown in Figure 3.12.

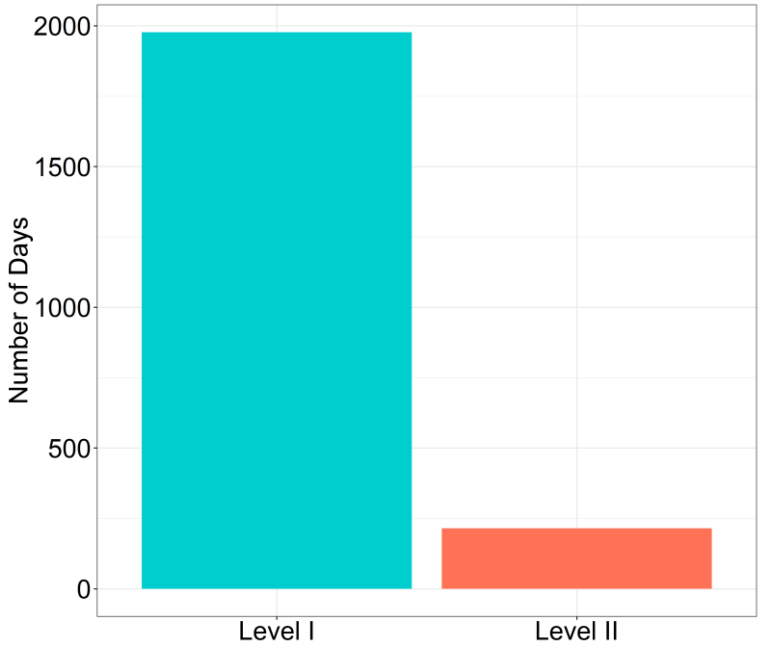


Figure 3.12 Two levels of daily outages

In order to intuitively demonstrate the weather’s impact on the two different outage levels, the PCA method is utilized to reduce the multi-dimensional weather conditions and visualize the two levels on a 2-dimensional plot.

PCA is an orthogonal transformation for dimensionality reduction [164], whose principle is to find a set of optimal base vectors **p** to represent the original data **X** in the original space as **Y** in the new space. In this research, the 8 weather features are regarded as 8 dimensions of the dataset in the original vector space. The covariance matrix of the data in the new space after dimensionality reduction is calculated as follows:

$$\mathrm{Cov}(\mathbf{Y}) = \frac{1}{m}\mathbf{Y}\mathbf{Y}^T = \frac{1}{m}(\mathbf{pX})(\mathbf{pX})^T = \mathbf{p}\left(\frac{1}{m}\mathbf{X}\mathbf{X}^T\right)\mathbf{p}^T \tag{3.12}$$

where m is the number of records in our weather dataset.

With an optimal set of base vectors **p**, the covariance of matrix **Y** is supposed to be a diagonal matrix with the covariance of weather features in the new space as 0 and the self-variance as large as possible. It is a classic mathematical problem as orthogonalization of real symmetric matrices. The eigenvalues of the covariance matrix in the original space indicates the percentage of the base vectors (i.e., principal components) in the new space as shown in Figure 3.13.

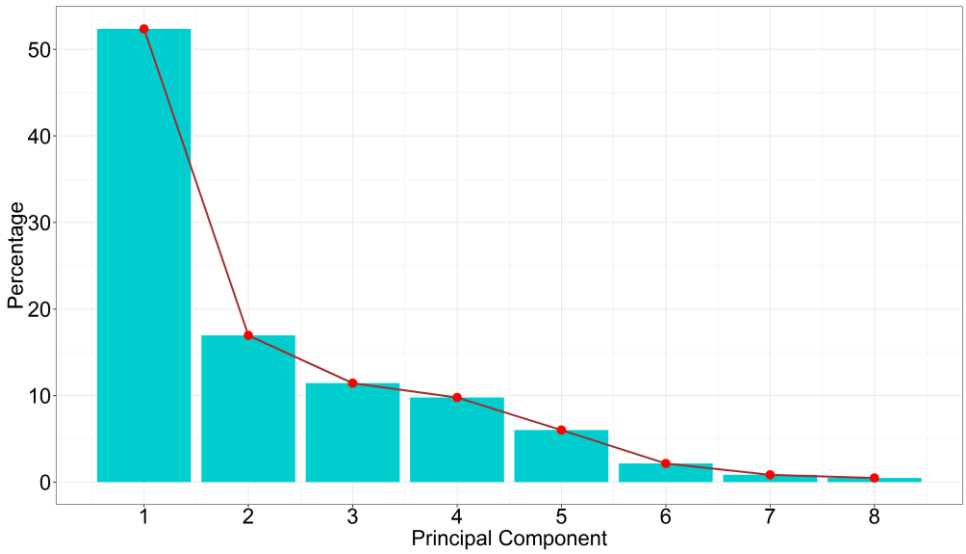


Figure 3.13 Percentage of principal component

According to the figure above, around 70% information could be represented with the first two largest principal components. Therefore, the daily records during 6 years could be visualized as a 2-dimensional graph in Figure 3.14 without losing most of the information.

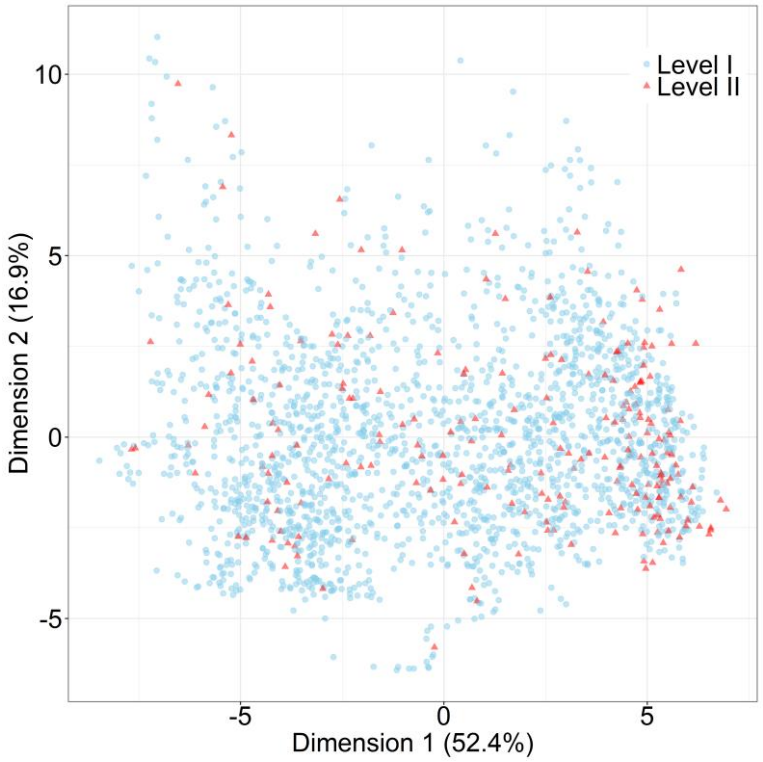


Figure 3.14 Visualization of two outage levels

As shown in Figure 3.14, the days defined as outage Level II are labeled in red and the days of Level I are in blue. The spread of red points in general overlaps the scatter of blue points, while the red points are more concentrated on the right part of the graph. Instead, the density of blue points on the right part is not much different to the left part of the graph. This indicates that the outage

Level I and Level II could happen in most of the weather conditions and consequently it may be difficult to find a hard boundary purely based on the weather conditions to determine the outage levels. However, the dense red points on the right part of the graph still show that the outage Level II has a higher probability to happen under some certain weather conditions. To reveal the weak relations between the weather conditions and outage levels, a data-driven model is to be built in the following part.

3.2.2 Classification

As discussed above, the outages in urban distribution network are divided into two levels according to the severity. Therefore, the task becomes a binary classification problem about the outage levels under given weather condition. In this study, the classification model is to be built based on SVM algorithm.

A. SVM algorithm

SVM is a classic learning algorithm which involves both structural risk minimization principle and statistical learning theory [165]. Given a dataset denoted as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where m is the size of dataset, $\boldsymbol{x}_i$ is a d-dimensional vector representing the features of the i-th sample and $y_i \in \{-1, +1\}$ as the corresponding class labels. The objective of a binary classification model is to find a mathematical function $h(x_i)$, which satisfies the following equation

$$y_i h(\mathbf{x}_i) = 1 \tag{3.13}$$

In particular for a linear binary classifier, the above equation can be re-written as below

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0 \tag{3.14}$$

where the classification model can be expressed as

$$h(\mathbf{x}_i) = sign(\mathbf{w}^T\mathbf{x}_i + b) \tag{3.15}$$

In an ideal linear binary classification model, the two classes are separated by a hyperplane described with a group of parameters $(\mathbf{w}^T, b)$. The principle of SVM algorithm is to widen the margin between the decision hyperplane and samples for maximizing the generalization capability of the model. Therefore, the mathematical expression of the margin between a data point $\mathbf{p} \in \mathbb{R}^d$ which is the closest one to the hyperplane $(\mathbf{w}^T\mathbf{x}_i + b) = 0$ is used as the objective function in SVM as below:

$$\max_{\boldsymbol{w},b} \{\min_p (\frac{1}{\|\mathbf{w}\|}\left|\mathbf{w}^T\mathbf{x}_p + b\right|)\} \tag{3.16}$$

As can be proven, the scaling of parameters in the hyperplane's expression dose not affect the optimal solution of equation (5). Hence, the basic format of the linear SVM could be expressed as

$$\min_{\boldsymbol{w},b} (\frac{1}{2}\mathbf{w}^T\mathbf{w}) \tag{3.17}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1 \tag{3.18}$$

In order to avoid the overfitting problem, a slack factor ξ_i is then introduced to allow a small portion of samples classified into the wrong classes via the model, whose expression is as follows:

$$\xi_i = \begin{cases} 0 & if\ y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1 \\ 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) & else \end{cases} \tag{3.19}$$

Then, the optimization problem in SVM could be formulated as:

$$\begin{cases} \min_{w,b,\xi} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i\right) \\ s.t.\ y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1 - \xi_i \text{ and } \xi_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \tag{3.20}$$

where C is the penalty parameter that allows a small portion of misclassification error during the maximization of margin.

According to the optimization theory, the above problem is equivalent to its dual formulation by introducing the Lagrange multiplier α as shown below:

$$\begin{cases} \min_{\alpha} \left(\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\, \mathbf{x}_i^T\mathbf{x}_j^T - \sum_{j=1}^{m}\alpha_j\right) \\ s.t.\ \sum_{j=1}^{m}\alpha_j y_j = 0 \text{ and } C \geq \alpha_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \tag{3.21}$$

The modeling of linear SVM finally becomes an optimization problem for finding a set of optimal parameters α_i $(i = 1, \dots, m)$ which satisfy the equation (3.21). With the obtained values of α, the parameters of hyperplane and could be calculated. For those samples which satisfy the equation (3.22) are regarded as support vectors.

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 \tag{3.22}$$

More generally, for the cases which could not be linearly separated, a kernel function will be introduced to map the original samples to a higher dimensional feature space. One of the typical kernel functions is as in equation (3.23)

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2} \quad \gamma > 0 \tag{3.23}$$

B. Over-sampling technique

As previouy discussed, the two outage levels could be taken as two classes in the binary classification problem. However, there exists a severe imbalance between Level I and Level II. In our study, the synthetic minority over-sampling technique (SMOTE) [163] is utilized to over-sample the minority class (i.e., Level II in our case) for balancing the original samples. The synthetic samples are generated alone the line segments joining any of the *k* nearest neighbors in the minority class.

The procedures of SMOTE are as follows: For each sample $\boldsymbol{x}_\mathrm{i}$ in the minority class, the *k* nearest neighbors in the same class is located based on the Euclidean distance. Then one of these neighbors $\boldsymbol{x}_\mathrm{j}$ is randomly selected. The synthetic

sample $\boldsymbol{x}_{\text{syn}}$ is generated on the line segment of every dimension between the original sample and the selected neighbor according to equation (3.24)

$$\mathbf{x}_{syn} = \mathbf{x}_i + \delta(\mathbf{x}_i - \mathbf{x}_j) \tag{3.24}$$

where δ is a random value within [0,1].

3.2.3 Case study and Results

The SVM-based binary classifier is modeled with the weather conditions and corresponding outage records introduced in Section II. The performance of the model is evaluated upon the test dataset which has not been used for training the model. The initial output is the probabilities of the sample belonging to different classes.

With the help of ROC curve as shown in Figure 3.15, the balance between the True Positive Rate (TPR) and False Positive Rate (FPR) could be achieved with the area under curve as 0.651. Finally, 83% of the samples in Level I and 43% samples in Level II are successfully identified.

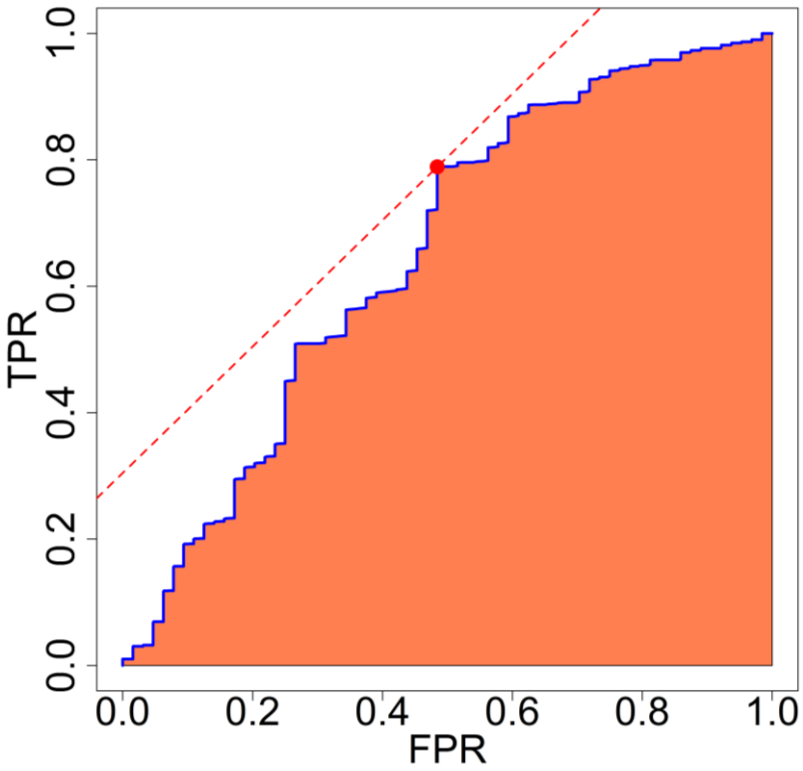


Figure 3.15 ROC Curve of the SVM Classification Model

3.3 Summary

In this chapter, two applications of data analytics on the outages in distribution system have been presented. As the terminal of power grid, the distribution system directly connects the customers and is sensitive to the consumers' satisfaction. An accurate prediction of the outage number in the following months provides a useful reference for the maintenance planning.

Moreover, it is also an instructive indicator for the investment in the construction of a stronger power grid in a larger time framework.

An improved grey model is firstly introduced to forecast the number of outages at the annual or monthly level. The parameters for building the traditional grey model GM(1,1) is improved with the optimization techniques. With the MAPE as an objective function, the optimized parameter successfully minimizes the error between the prediction results and real values. Since the monthly outage number is a seasonal data sequence, the rolling mechanism is introduced to improve the ability of capturing the change of trend.

As for the daily outages, a data-driven model is built to classify the outage levels according to the number of failure records per day. The results showed a positive but weak relation between the daily outages and weather features considered in this paper. The future work will focus more on the collection of effective features and improve the model for predictive maintenance.

Chapter 4

Anomaly power consumption detection from incomplete records

4.1 Introduction

With the rapid development of information and communication technology, advanced metering infrastructures have been widely applied in power system at the customers' level. In US, the installation of smart meters has seen an increase since 2007. By the end of 2016, 72 million smart meters have been installed among over 55% of US houses [166]. The percentage of electricity customers in EU28+2 with smart meters are expected to reach 71% by 2023 [167]. In particular, the first generation of smart meters has started in Italy since 2001 and covered 95% of 36 million customers by the end of 2011 [168].

There have been already many researches focusing on the technical and optimal operations of distribution system [169-170]. While with the wide application of smart meters, the data analytics methods start to support the secure and economic operation of power systems. One of the practical applications with smart meters in power system is to detect the abnormal behaviors of customers, which is long of concern among utilities [171-172]. The hourly recorded energy consumption of building system has been analyzed in [173] with the classification and regression tree algorithm. The abnormal energy consumption is then detected by the generalized extreme studentized deviate algorithm. In the photovoltaic system, the fault degree is detected and evaluated with the local outlier factor (LOF) based algorithm in [174], which distinguishes the abnormal data with specific mathematical characteristics.

The recorded data regarding the customer consumption pattern may be not complete: in this chapter, a possible solution for this situation is proposed, and the incomplete data are used for investigating the electricity consumption of individual customers. Since civilian customers' electricity consumption is normally affected by the weather conditions [175], a proper description of weather

sensitivity among customers based on only the trusted part of records could be used for further analysis, including abnormal behavior detection. As one of the most widely used methods for sensitivity analysis, regression model is used to quantify the impact of model-input variables thanks to the fast computing and easy interpretation [176]. For example, a method of Bayesian Additive Regression Trees (BART) is introduced and utilized in [177] to capture the climate sensitivity among electric power consumption data. Furthermore, the abnormal electricity consumption patterns among civilian customers can be detected from the incomplete dataset.

4.2 Structure of the anomaly detection method

The methodology proposed in this paper is summarized in Fig. 4.1 and is composed of three main steps.

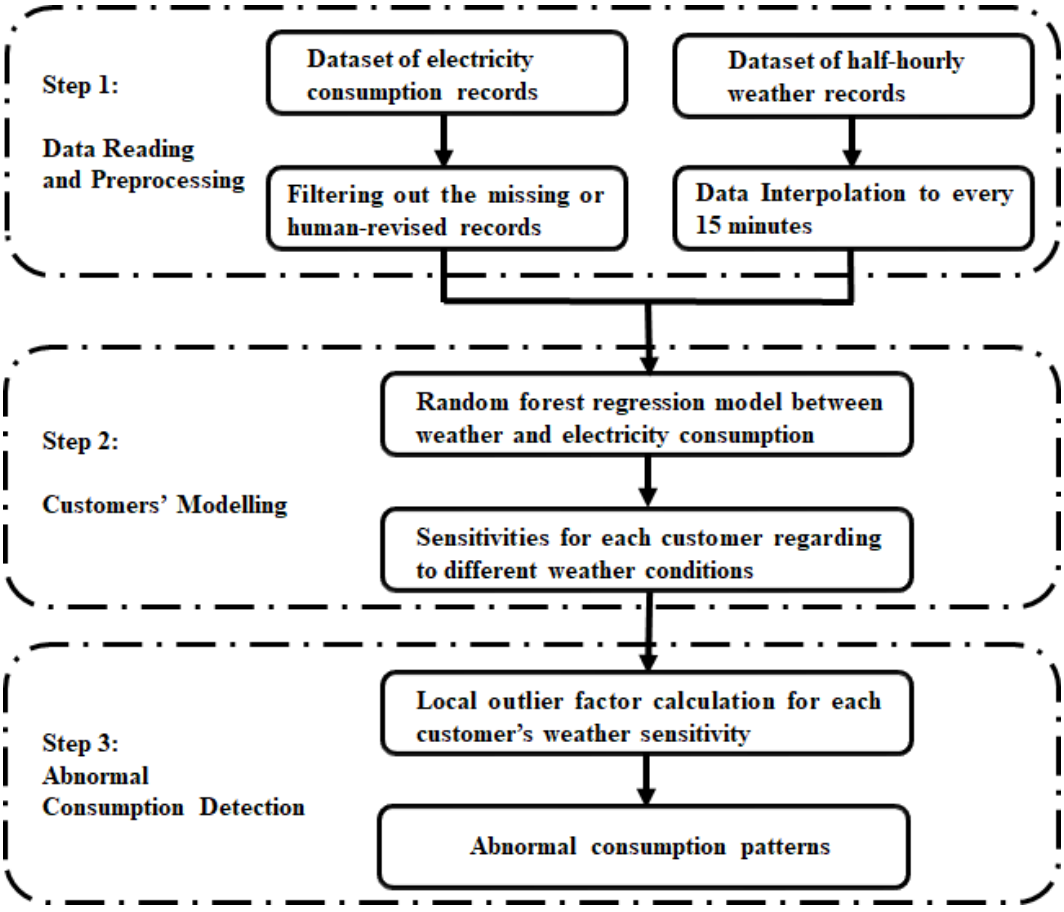


Figure 4.1 Scheme of anomaly detection from electricity consumption patterns

- Data reading and pre-processing: the data refer to civilian consumption and weather data in year 2015.
- Customers' modelling: this step focuses on the building of the regression model for revealing the relation between the weather conditions and the individual customer's consumption. The relatively importance of the different weather features considered in the model has been obtained through a sensitivity analysis.

- Abnormal consumption detection: thanks to the model obtained in Step 2, it is possible to detect the customers whose sensitivities are diverging with respect the majority of the other customers.

4.3 Data description

In this section, the datasets used in this paper is briefly introduced, including the civilian customers' electricity consumption records collected from the local Distribution System Operator (DSO) in Turin (Italy) and the weather condition records during the same period.

4.3.1 Electricity data

The records of electricity consumption in every 15 minutes for civilian customers are collected by the DSO. All the customers' information is anonymous with only a specific reference number for identification as in Fig. 4.2.

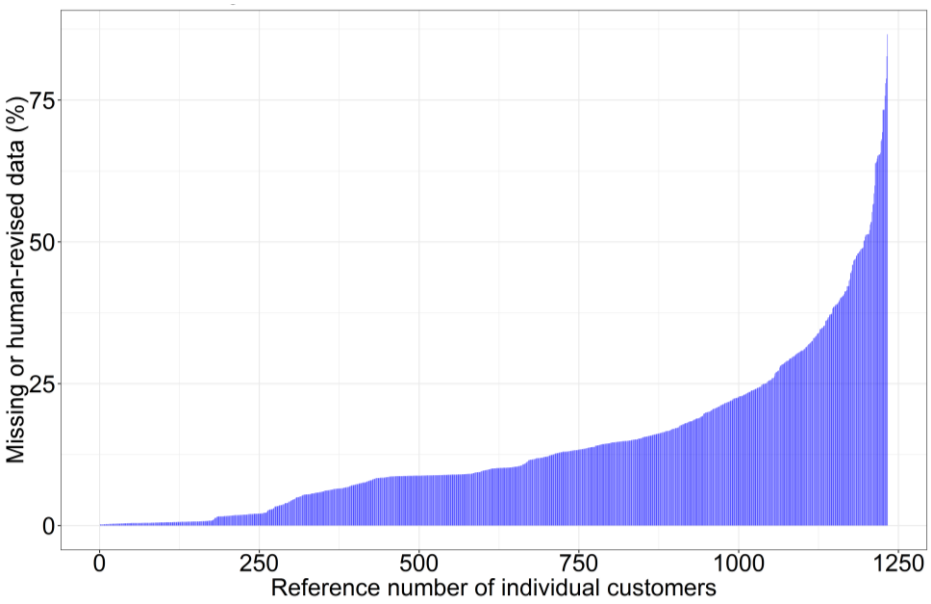


Figure 4.2 Quality of the civilian customers' electricity consumption records

As mentioned above, there are some missing and human-revised records in this dataset, which cannot be trusted in the regression model for sensitivity analysis. The percentage of untrusted records during the whole year for each customer is calculated and shown in Figure 4.2. The civilian customers are sorted according to their percentage of untrusted records in an ascending order at the horizontal coordinate axis. Although all the civilian customers have untrusted records, over 95% of them have less than 50% missing or human-revised data. One typical electricity consumption curve of a customer during an entire week is presented in Figure 4.3. The missing part of the curve on Wednesday is due to the missing records in the original dataset. The red part is also from the dataset while with a special note as "revised". Since this part is far away from the typical consumption patterns, we could not use them for customers' behavior analysis.

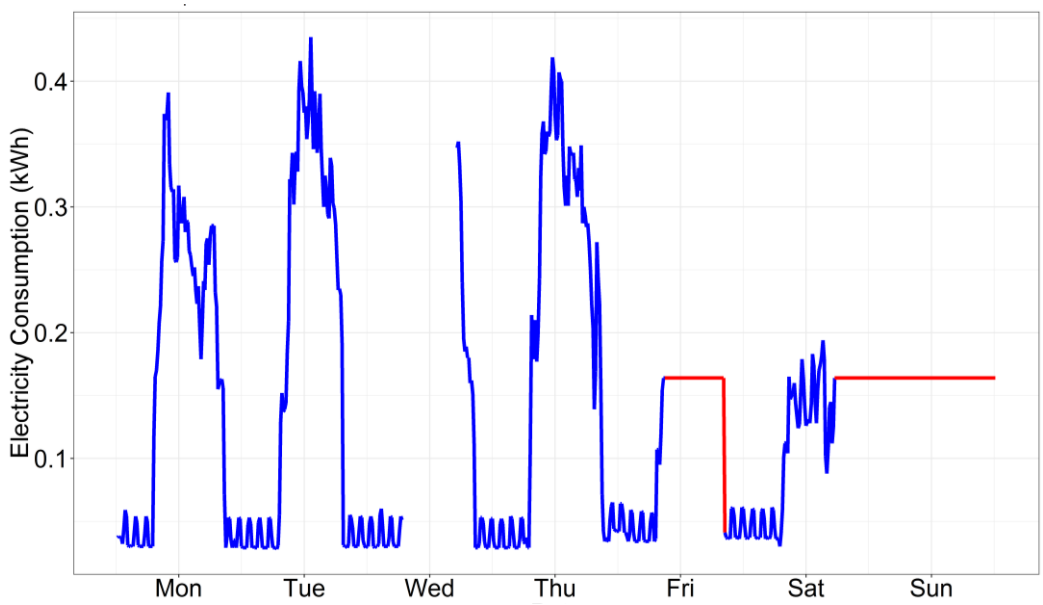
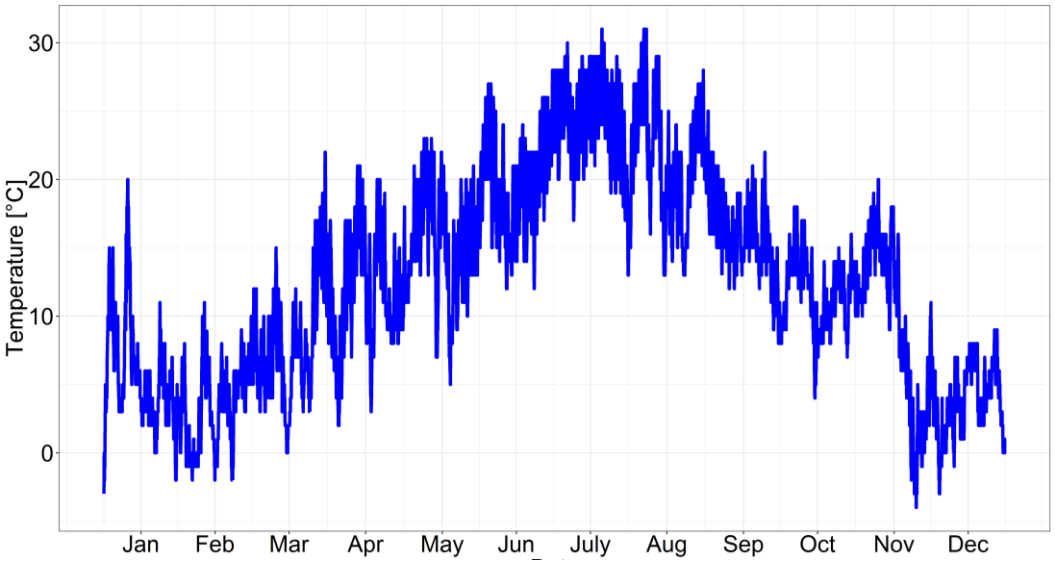


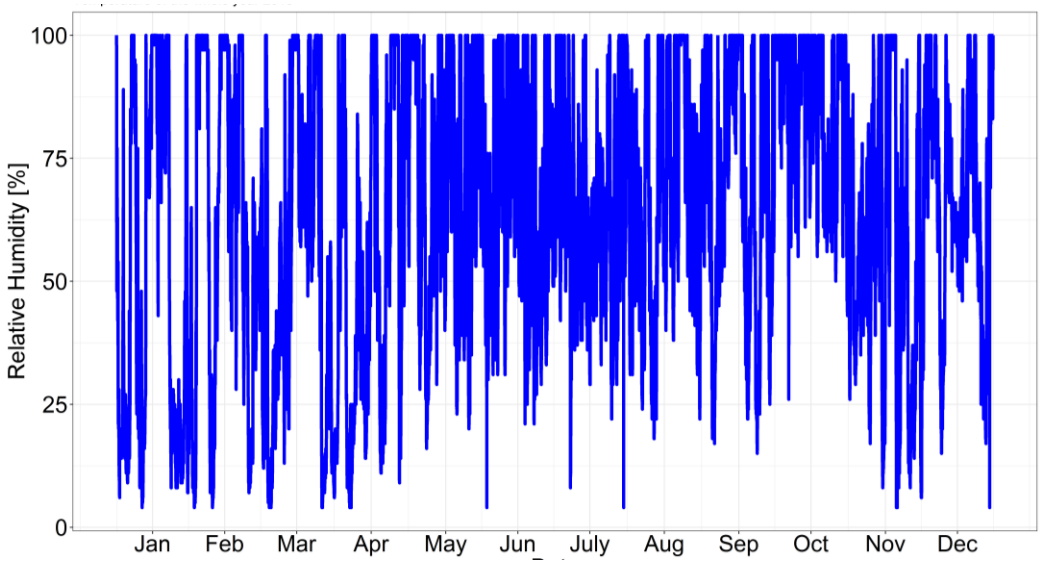
Figure 4.3 A typical incomplete electricity consumption curve for one week

4.3.2 Weather data

The weather information covers all the year as shown in Fig. 4.4.



(a) Temperature



(b) Relative humidity

Figure 4.4 A typical electricity consumption curve for one week

The weather data in this research include temperature, humidity, dew point, wind speed, atmospheric pressure and visibility at the sampling rate of half an hour. In coordination with the electricity consumption data, the linear interpolation is adopted in this work to fill the missing weather conditions in the middle of half an hour.

4.4 Customers' sensitivity

Sensitivity analysis is an important tool to identify the key variables affecting the energy use from observational study [177]. The impact of climate factors on the energy analysis in buildings has already been conducted through the sensitivity analysis in some researches [174, 178]. In this paper, a regression model-based sensitivity is applied in the datasets and reveals the underlying importance of different weather features to the use of electricity.

4.4.1 Regression model

Regression model-based method is a typical approach for the sensitivity analysis thanks to its clear meaning and easy interpretation. The critical process is to build an accurate regression model between the input variables and the output. In our case, each individual customer will be described with a specific regression model for the different electricity consumption characteristics. The six weather features at each quarter of an hour are the input variables of the model while the corresponding electricity consumption quantity during those 15 minutes is taken as the output. Since there is no need to build the regression model based on the sorted data according to sampling time, the missing and revised records at different periods in the dataset is no more a limitation for the application of the incomplete data. For a complete electricity consumption dataset, there are supposed to be 35040 records for each civilian customer. In data preparation, all

the missing and human-revised records among the electricity consumption dataset are filtered out at first. In this work, the number of customers N_{cust} = 1150，corresponding to the customers with more than 50% of the total records (i.e., ≥17520) valid.

Due to the complex impact of weather conditions on the use of electricity in civilian customers, it is difficult to describe the relations with traditional linear and non-linear regression models [177]. Random forest regression model, as a powerful and classic machine learning technique, is a promising solution to this kind of analysis.

Random forest is a machine learning technique stems from the concept of Classification and Regression Tree (CART) [179]. With the bootstrap resampling of data, several subsets of samples are used to train different CARTs, which is known as bagging regression trees. Compared to the other complex machine learning techniques like neural networks, the computation of CART decreases a lot by dichotomizing the input variables. The random forest algorithm in Figure 4.5 further introduces the randomness in subset of features used for training independent trees as weak classifiers or regressors. Finally, the results of de-correlated CARTs are collected as the final output with a significant improvement of accuracy.

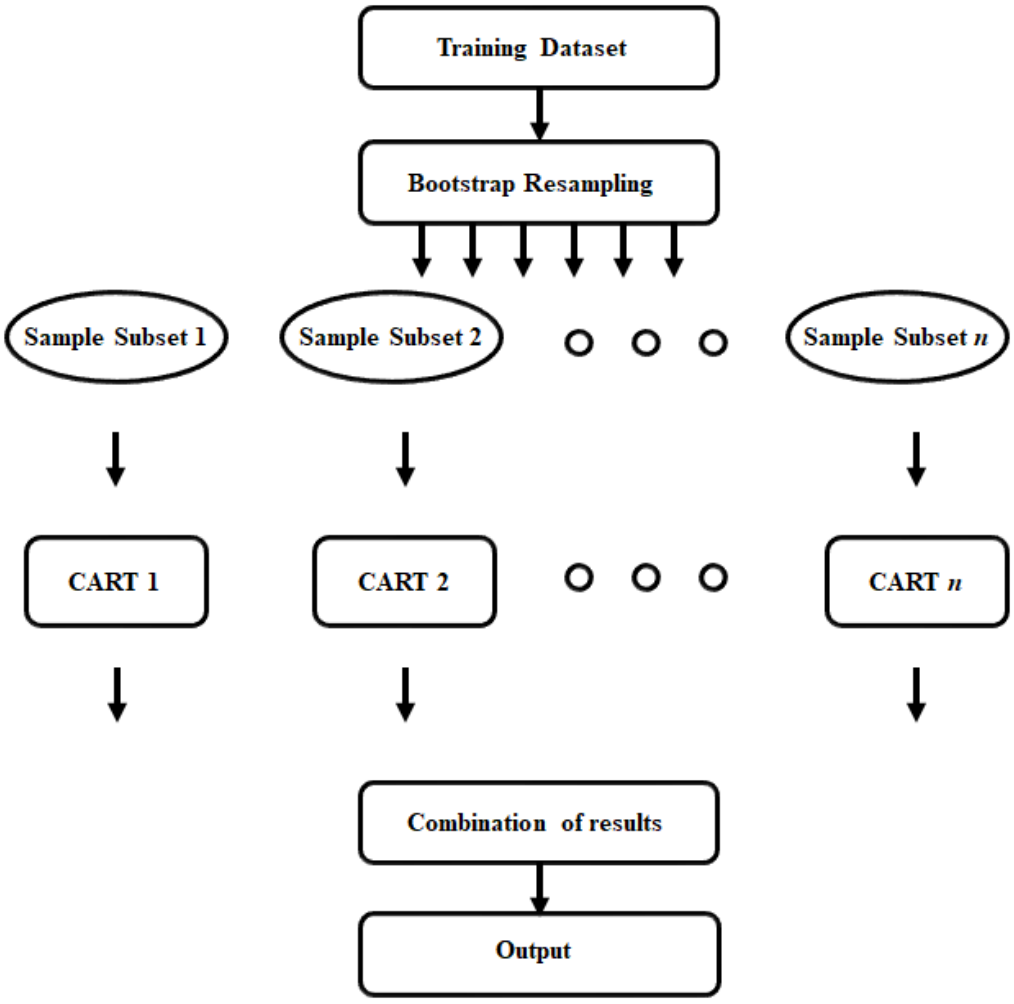


Figure 4.5 Typical structure of random forest

To construct the structure of random forest, the critical parameter needs to be determined is the number of CARTs. Typically, the more trees involved in the random forest, the more accurate the output would be compared to the customer's real consumption, while at a higher cost of time consuming.

In random forest algorithm, one part of the training data is reserved for each tree as the out-of-bag (OOB) samples in the bootstrap resampling step. The accuracy of regression model can be tested on OOB samples. One of the typical evaluation indices is the mean-square error (MSE) in (4.1):

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)^2 \tag{4.1}$$

where m is the size of OOB samples, y_i is the real value of the i-th sample and $\hat{y}_i$ is the output of regression model. The effect of the number of trees could be evaluated with the performance of random forest model as shown in Fig. 4.6.

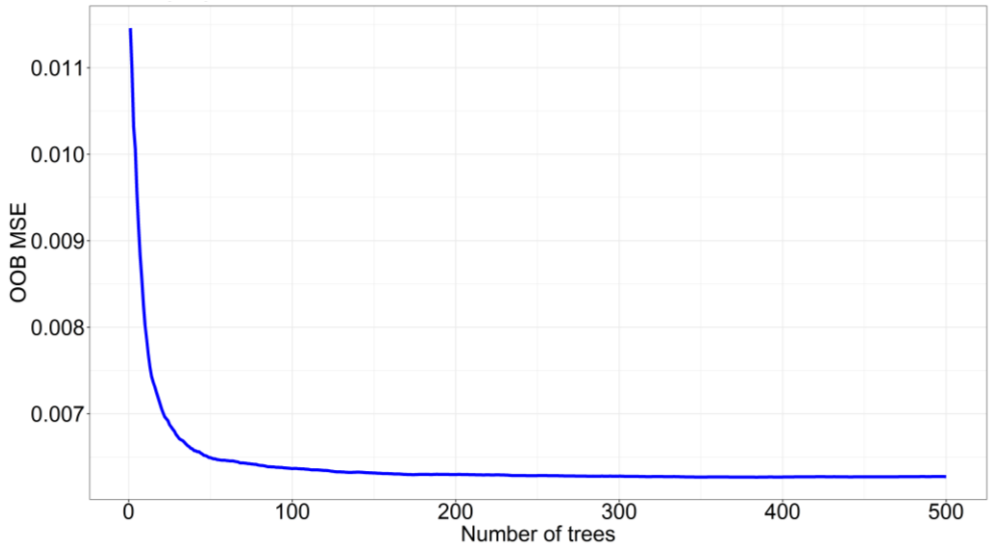


Figure 4.6 OOB MSE with different number of trees

In Fig. 4.6, a typical customer's electricity consumption records are analyzed by the random forest regression model with different number of trees. In this case, the performance of the regression model becomes stable after 100 trees.

4.4.2 Sensitivity analysis

In classic linear or nonlinear parameterized regression models, the impact of key input variables on the output are determined with the standardized regression coefficients [177]. However, as a non-parameterized model, there is no predetermined form in the random forest regression. Alternatively, the sensitivity of electricity consumptions to the input variables could be evaluated with the importance of variables in building the random forest regression model.

In the modeling of random forest for a specific customer, the variable's importance is also evaluated with the OOB samples by three steps:

- Calculate the OOB MSE of the j-th CART in random forest, denoted as E_j
- Introduce noise randomly to the i-th variable, and then calculate the OOB MSE for the j-th CART, denoted as E_{ij}^*
- Calculate the importance of the i-th variable by considering all the n CARTs as in (4.2)

$$f_i = \sum_{j=1}^{n} \left(E_j - E_{ij}^* \right)^2 \tag{4.2}$$

If the accuracy (4.1) of random forest decreases a lot after introducing the noise into the i-th variable, the two OOB MSEs E_j and E_{ij}^* would be different to a large extent, which leads to a high value of f_i. This means that the i-th variable has a high impact on the performance of the regression model.

By repeating the evaluation steps over all the variables, the importance of weather features to a specific customer could be determined, which is also recognized as the sensitivity of electricity consumption to different variables in our case. As an example, for the customer analyzed in Fig. 4.7, the electricity consumption is more sensitive to the temperature, dew point and atmospheric pressure than the relative humidity.

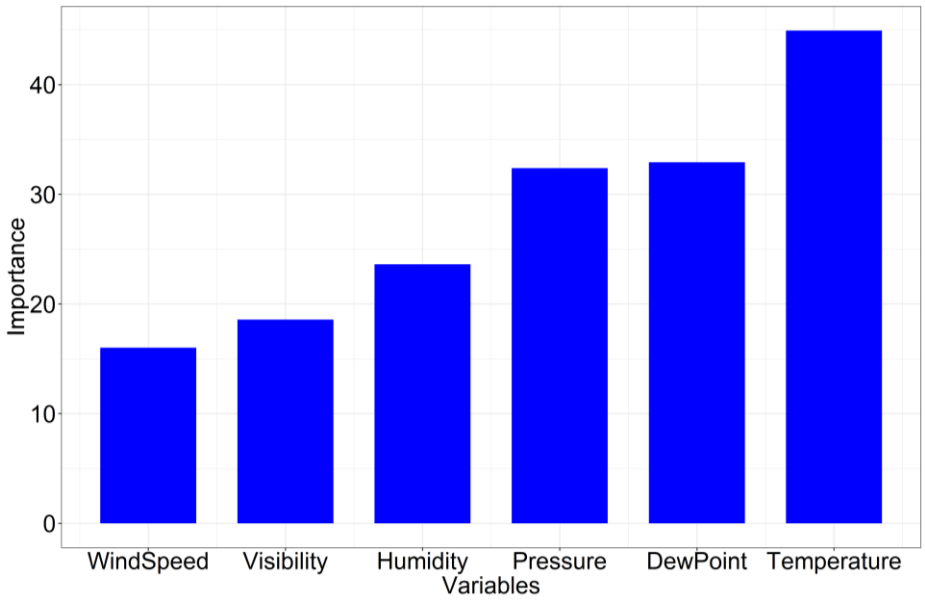


Figure 4.7 Variable importance of weather features to the electricity consumption

4.5 Anomaly detection

Even though the collected dataset is incomplete as demonstrated previously, it still contains valuable information and needs efforts to be explored. In our study, all the civilian customers could build their regression models from only the trusted part of the consumption records with the weather conditions at the same sampling frequency. Based on these regression models, the civilian customers' electricity

consumption sensitivity to various weather conditions could be obtained as a complete dataset for describing their characteristics.

Different civilian customers may behave largely different to weather conditions. With the same variation of weather conditions, their consumptions may change accordingly to different degrees.

The sensitivity analysis quantifies such change in a relative value for each customer instead of the absolute value. In general, almost all the civilian customers should react more or less in regard to the environment. Therefore, the abnormal electricity consumption detection could be accomplished by locating the *outliers in the customers' sensitivities* to weather features.

4.5.1 LOF-based outlier detection

Outlier detection is an important data mining method. Without the prior knowledge of normal customers' sensitivities, the distribution-based outlier detection is not much practical in this case. LOF, a density-based method proposed in [180], shows good properties by assigning to each customer a degree of being an outlier. To get the local outlier factor, the concepts for the density evaluation of object p is introduced as below.

In the dataset X, the distance between o and all the other objects p are calculated and sorted in an ascending order. The k-distance of object o is defined as the k-th distance between o and all the other objects, denoted as $dist_k(o)$. The k-distance neighborhood of o is the objects within the $dist_k(o)$. Then, the reach-distance of p is defined as $reach\text{-}dist_k(p,o)$ in (4.3).

$$reach-dist_k(p,o)=\max\left\{dist_k(o),d(p,o)\right\} \tag{4.3}$$

where $d(p,o)$ refers to the distance between object p and o.

The local reachability density of p is defined as below in (4.4).

$$lrd_k(p)=\frac{k}{\sum_{o\in N_k(p)} reach-dist_k(p,o)} \tag{4}$$

Finally, the LOF index $LI_k(p)$ referring to the k-th distance of point p is defined in (5)

$$LI_k(p)=\frac{1}{k}\sum_{o\in N_k(p)}\frac{lrd_k(o)}{lrd_k(p)} \tag{5}$$

The definition of $LI_k(p)$ is the ratio of the neighbors' average local reachability of point p to its own local reachability. If the object under test is not an outlier, this value should be very close to 1. On the contrary, for the outliers, such ratio is far from 1 due to the large distance to its neighborhoods. The LOF index $LI_{50}(p)$, with $p=1,\ldots,N_{cust}$, referring to the 50-distance (i.e., $k = 50$) of all the customers from our sensitivity dataset is calculated as shown in Figure 4.8.

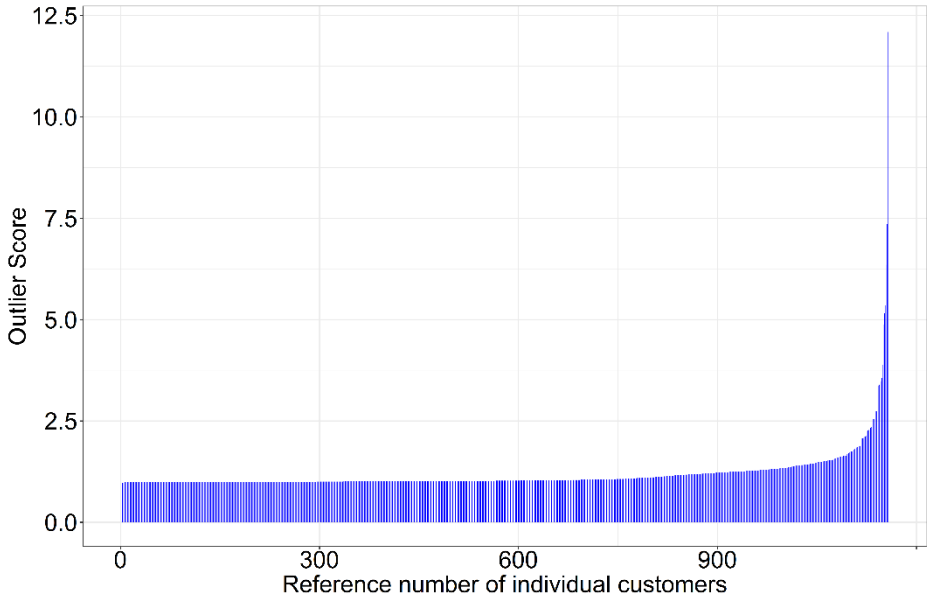


Figure 4.8 LOF of all the customers' weather sensitivity values

As can be seen from Figure 4.8, most of the customers p=1,…, N_{cust}, have a value of $LI_{50}(p)$ around 1. Only 44 customers are detected out with a $LI_{50}(p)$ larger than 2. Most of these selected customers have large period of the year with a pretty flat electricity consumption curve. It may be caused by the frozen warehouses or some apartments without any people living in for several months while with some appliances on. A further detailed investigation could be conducted on those customers for improving energy efficiencies or detecting non-technical losses.

4.5.2 Entropy-based outlier detection

As mentioned above, one of the challenges for outlier detection is the incompleteness of electricity consumption records, which makes it impossible to compare the xcurves among different customers. However, for each customer, the electricity consumption at each hour of the day is relatively stable. Some missing values will not affect the hourly statistic results. Therefore, we could focus on the hourly statistics of an individual customer, instead of the time series records.

The "stability" of the consumption inside one hour over all the year's records could be mathematically described with the concept of Entropy:

$$E(X) = -\sum_{x \in X} P(x) log P(x) \quad (6)$$

In our study, the range of consumption of an individual customer in each hour could be firstly discreted into 10 intervals. The probability of electricity consumption at each interval of that hour could be calculated as $P(x)$ in (6). Then, the entropy of electricity consumption in that hour could be obtained with equation (6). Finally, the 24 entropies over all the hours of a day in the year of record could be calculated. The maximum and minimum values of hourly electricity consumption in 24 hours of a day over the year are also collected as important features of a customer. The characteristic of an individual customer is

now described with these 72 features. If the degree of fluctuation in the hourly electricity consumption of a particular customer is significantly different from that of most of the others, the behavior of that customer is detected as abnormal electricity consumption. In order to identify the outliers based on the 72 features, the density-based spatial clusteirng of applications with noise (DBSCAN) is utilized for analysis.

Unlike typical clustering algorithms, the DBSCAN does not require a pre-defined number of clusters, but shows a good performance in identifying outliers. Instead, two important parameters are needed: the radius of neighborhood around a given data point, *epsilon*; and the minimum number of neighbors within the epsilon radius, *MinPts*. With these two parameters, three types of data points could be identified:

- *Core point*: a data point with a neighbour count greater than or equal to the *MinPts*.
- *Border point*: a data point with a neighbour count less than *MinPts*, but belonging to the neighbourhood of some other data point.
- *Outlier*: a data point neither a core point nor a border point.

The algorithm of DBSCAN can be summarized as below:

- *Step1*: For each data point, compute the distances from it to the other points and count the number of neighbours within the radius *MinPts*.
- *Step2*: For each core point, if it has not been assigned to a cluster, create a new cluster. The neighbours of the core points within the radius of the starting core point are assigned to the same cluster.
- *Step3*: Iterate through the remaining points in the dataset.

With this method, when we choose both *epsilon* and *MinPt* as 5, there are 41 customers identified as abnormal ones according to the whole year's records, among which 15 customers have been also identified with the weather sensitivity outlier detection method.

4.6 Summary

This chapter proposed a data-driven method for evaluating the weather impact on the behavior of civilian customers' electricity consumption. This work could be taken as the base for non-technical loss detection in distribution network. The first challenge in our work is to deal with the incomplete records in the dataset collected from the first-generation smart meters. Investigations on the consumption curves at the same time scale become difficult due to the randomly missing and human-revised data among different customers. Instead, we used an indirect method to study the sensitivities of individual customers only based on the trusted part of data.

As a non-parameterized model, random forest regression in this paper is used to describe the relations between electricity consumption and weather conditions.

In this model, the sensitivities are evaluated by the importance of different features. All the sensitivities are trustworthy only when the regression could accurately reveal the behavior of customers.

By building individual regression models for all the customers, their sensitivities could be obtained as a new dataset for abnormal behavior detection. Since the models are built in a customized way, the absolute quantity of power variation to different weather conditions will not affect much in the sensitivities. Neither increase nor decrease of the power demand will be regarded as the same reaction to the change of environment. However, only those customers who hardly have any reactions to the environment or changed their behavior largely are detected as the abnormal cases. Finally, limited number of abnormal customers are successfully detected out for an incomplete dataset. This information could be used by the power utilities for a further investigation.

Chapter 5

Monte Carlo simulation-based analysis on high-impact low-probability events in distribution system

5.1 Introduction

As the backbone of modern industrialized society and economics, the uninterrupted power system supplies play a significant role in people's daily life [181]. However, with the deterioration of global climate, electrical power critical infrastructures are exposing to a harsher environment all over the world [182]. In recent years, many power outages triggered by the natural hazards have been witnessed. For example, over 500 000 Long Island Power Authority customers lost power in 2011 after being hit by the hurricane Irene [183]. It took 8 days before the restoration of 99% of the company's customers. The total number of customers affected by this hurricane is more than 4.3 million people on the east coast of the US.

Since people's life could be hardly independent from the stable power supply, the utilities are under huge pressure for providing a reliable service under different threats from the environment. In distribution network, the consequences to customers of long-time blackouts include disconnection of internet at house, halt of work in the office, employment of the backup power sources, and so on. As to the DSO, they have to pay not only the cost of repairment in the power equipment, but also the loss of customers. Therefore, the imperative demand for a more reliable power system is from both sides of the power grid, which drives the utilities to invest more in the infrastructure enhancement to the adverse environmental conditions. It has been reported that in Europe, the share of investments in distribution network over the power grids is supposed to grow

continuously from 66% in 202 to 80 by 2050 [184]. With the percentage of people living in the urban area approaching 68% of the total population around the world in 2050 [185], the urban power distribution network is deeply involved in most people's life acting as the "final mile" in the delivery of electricity.

There are some researches focus on the weather's impact to the power system. For example, the effects of catastrophic weather on the reliability of power system is evaluated with a fuzzy clustering method in [186]. The cascading failures in power grid caused by natural hazards is analyzed with an extreme weather stochastic model in [187]. In Italy, with the heat waves in urban areas, increasing faults have been recorded in distribution network in a relatively concentrated period [189]. These dense occurrences of faults bring a severe threat to the secure operation of the urban distribution system. Because as a meshed network operating in radial, the single fault on the feeder is possible to be isolated without losing any customers for a long time. However, the heat wave could cause the second fault happening on the same feeder in a short time before the first one is repaired, which will lead to a severe blackout for the customers between the two fault locations. This is a typical extreme weather-related power interruption with high impact and low probability.

In addition to the reliable operation of power system under common weather conditions, the ability to withstand extraordinary and high-impact low-probability events is also needed in the planning and operation of modern power system [182]. Because the rare events could cause huge damage to the fundamental infrastructure and have a tremendous social impact. Therefore, the concept of resilience becomes an emerging topic in power system [190]. A probabilistic methodology is proposed in [191] to evaluate the adaptation measures to increase the resilience of power system to natural disaster. This method is also capable to deal with the multi-hazard and multi-risk analysis power system resilience. A multi-phase resilience assessment framework is developed in [192], which is used to analyze the natural threat on critical infrastructures. Different strategies to boost the resilience of power systems are also discussed with the multi-phase adaptation cases.

From the planning point of view [183], the Cost-benefit Analysis (CBA) is an important approach to the resilience problem. However, the low probability of extreme weathers brings a lot of difficulties to the CBA. Moreover, the uncertainties of occurrence of natural disasters need an appropriate methodology to be mathematically summarized and explained. In this chapter, the heat waves in an urban area of Italy are taken as the high-impact low-probability events to the local distribution system. In the last few years, the temperature in summer is becoming higher with a visual growth of power interruptions. Compared with the single fault in medium voltage feeders, the repetitive faults in the same feeder result in long-time power unavailability to both residential and industrial customers and cause huge economic losses. The occurrence probability of heat waves is discussed with statistics based on the weather records over 10 years, which is used to indicate the probability of critical years and non-critical years.

Then, a Monte Carlo simulation model is established based on the probability of extreme weather in the urban distribution system.

5.2 Heat wave and its impact

A better understanding of the critical weather conditions threating to the security of the power grid is the basis of our research. In general, the extreme weather refers to the severe conditions which does not appear often over a long period, such as hurricanes, droughts or heat waves. These events have a significant impact on the operation of electrical network due to their destroying power to infrastructures. In the meantime, long-time wide area electricity interruptions will disturb the normal people's life, business activities and even cause emergencies. Moreover, the frequency of extreme weather is increasing compared with the historical records with its consequences becoming more severe, which is a crucial problem to network operators around the world.

As a stress factor to the transmission and distribution grid, heat waves are attracting engineers' attentions to the reliability and resilience of the system. As a physical issue, the high temperatures brought by the heat waves will decrease the limit of the transfer capability of power cables, weaken the insulation and increase the energy losses [188]. According to the study of climate changes in [185], there is a gradual increase in frequency and severity of extreme heat waves in European metropolitan areas. In addition to the physical damage directly brought by the heat waves, the surge of power demand because of the extensive use of air-conditioners and electric fans increases the temperature of power cables due to large current flowing through and challenges the flexibility of power supply.

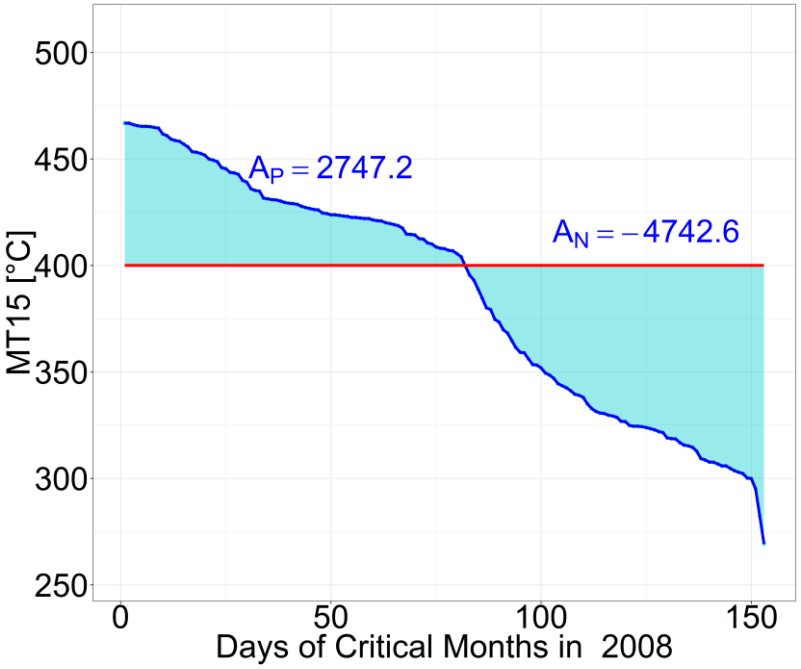
5.2.1 Occurrence of extreme heat waves

In order to better illustrate the character of heat waves and their occurrence, the weather records of Turin, the third largest city in Italy, is taken for analysis. In recent years, some Italian cities have witnessed more interruptions in distribution network among summer, which is recognized as an outcome of the heat wave [189]. Since high temperature is the most distinctive characteristic of heat waves, the maximum temperature is used for determining the occurrence in our study.

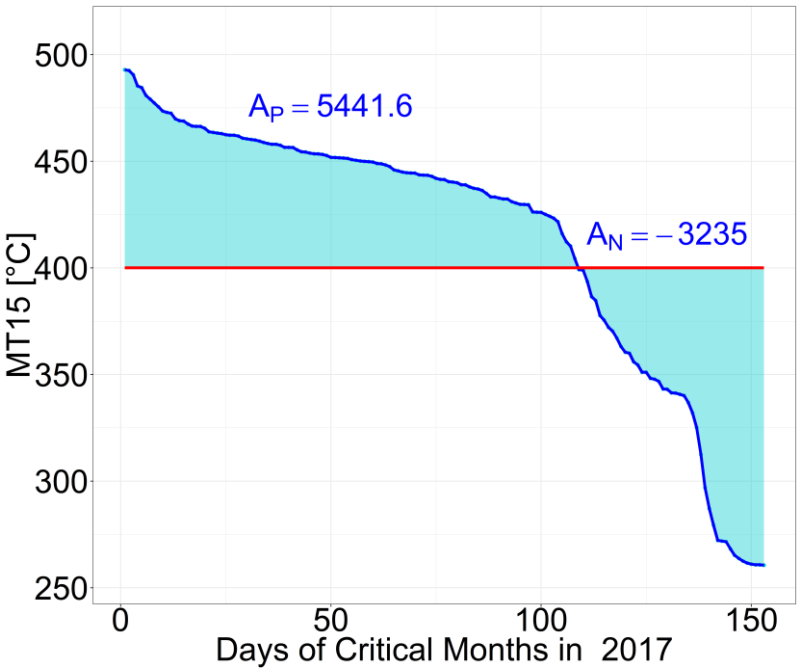
As the heat wave only occurs from May to September of the year, these months are labeled as critical months (CMs) in this paper. Correspondingly, the other months are taken as non-critical months (NCMs). It is obvious to point out that only the power interruptions during CMs are possibly related to the heat waves. Therefore, an efficient way to analyze the occurrence of heat wave is to split the year and focus only on the CMs.

Under practical conditions, the daily maximum temperature may fluctuate in consecutive days while the heat wave's impact on distribution network could be a cumulative consequence. To balance the peak temperature of the day and its continuous impact in a period, we sum the daily maximum temperature in 15 days as an index to evaluate the degree of hot weather. For each day, it has an index

based on the daily maximum temperature within 15 days, which is labeled as MT15. In CMs, there are 153 days (154 days for leap year) corresponding to the 153 MT15 values. By sorting the MT15 values in a descending order, a blue curve could be established as shown in Fig. 5.1.



(a) Area of MT15 in 2008



(b) Area of MT15 in 2017

Figure 5.1 Positive and negative areas of MT15 in 2008 and 2017, respectively

Although the curve of MT15 decreases in both of these two figures, their intercept on the vertical axis and the slopes are quite different. As can be seen

from Fig. 5.1, the highest value of MT15 in 2017 reaches almost 500, much larger than the value in 2008. For a clearer comparison, a red horizonal line is added at the value of 400. It is obvious that more days in 2017 remain above the red line. Moreover, the MT15 curve becomes steeper when the value is lower than 400 in 2017. Instead of counting only the number of days with MT15 values larger than 400, the light blue area of the curves above and below the red line are taken into analysis, which are labeled as AP and AN for positive and negative areas, respectively. The positive area AP of 2008 is 2747.2, almost half of the value in 2017. Typically, the positive area plays a more important role than the negative area because it reflects the characteristic of high temperature in a better way. If the positive areas of the two years are the same, the year with a less value of negative area means more critical in terms of the effect of heat wave. For example, the negative area $|A_N|$ of 2017 is 3235.0, which is less than the value in 2008. It is because that there are much fewer days with values less than 400 in CMs of 2017. With the records of 10 years from 2008 to 2017, the positive and negative areas of MT15 in each year with respect to the threshold of 400 could be calculated and shown in Fig. 5.2.

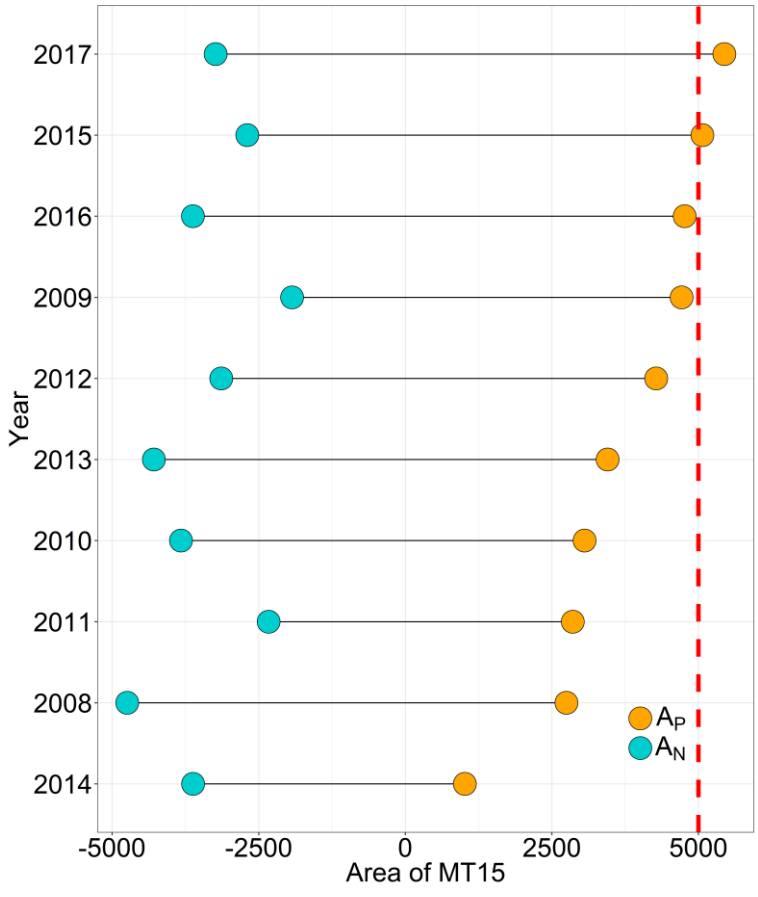


Figure 5.2 Positive and negative areas of MT15 from 2008 to 2017

In the figure above, the 10 years are sorted according to the positive areas displayed as yellow points. As can be seen in the figure above, in spite of the descending sort of positive areas, their corresponding negative areas have no such trend. Since the positive area is directly related to the severity of heat waves, the

years with positive area larger than 5000 are determined as years with extreme heat waves in our study. Only the year 2017 and 2015 over the vertical read dash line are deemed as the years when extreme heat waves came. Therefore, the occurrence probability of extreme heat waves in Turin could be determined as 2/10.

It could also be seen from the figure above that the worst heat waves happened in the last three years of the records, implying that the extreme weather is becoming more frequent and the threat of adverse weather conditions are emerging in the local urban area. In fact, the high temperature appears in every summer, which is not contrary to the design philosophy of infrastructure. However, the safety margin of design in the distribution network may lose its effectiveness with the appearance of extreme heat waves. This attracts a lot of interests on the topic of resilience in power system.

As for the daily maximum temperature in NCMs, it could be used to analyze the cold waves, which has limited impact on the distribution network and will not be discussed in this paper.

5.2.2 Impact of heat waves on the distribution system

According to the historical records from the local utility, the number of interruptions from 2008 to 2017 are shown in Fig. 5.3. Since only the interruptions in CMs are possibly caused by the heat waves, the annual records are divided into two parts: interruptions happened in CMs and NCMs. With the assumption that no interruptions in NCMs are related to the heat waves, the 10 years in the vertical axis are sorted merely according to the number of interruptions in CMs, which are labeled as yellow points. The records in NCMs are marked as points in light blue.

As can be seen from Fig. 5.3, although the number of interruptions in CMs are sorted in a descending order from top to bottom, the corresponding values in NCMs are unordered, which reveals the fact that the influence factors for interruptions in distribution network is complicated and the heat waves only account for part of the causes. Typical reasons for an interruption include the aging of equipment, natural hazards, manual error, and so on. The heat waves could only explain a small part of interruption records. However, by comparing the interruption records in the same period (like CMs) every year, it is possible to discover the potential impact of a specific phenomenon to the security of distribution network. For example, there are only two years from 2008 to 2017 deemed as the year with extreme heat waves, which are exactly the top two years in Fig. 3 due to the extraordinary large number of interruptions in CMs. Therefore, it is reasonable to assume that the increase of interruptions in 2015 and 2017 is highly related to the occurrence of extreme heat waves.

In the figure, there are two years (2008 & 2010) with the same number of interruptions in CMs and NCMs, whose yellow points are overlapped by the light blue points. In fact, the number of days in CMs are 153 (or 154 in leap years), accounting for only 42% of the whole year. If there is no seasonal effect on the

security of power system, more interruptions are supposed to be recorded in NCMs. However, throughout the ten years' records, the number of interruptions in CMs is larger or equal to that in NCMs for most of the years. It is consistent with our perception because the distribution network suffers from more challenges due to the high temperature. Even for the two years (2011 & 2014) when less faults were witnessed in CMs, the average number of faults per day is still much higher than that in NCMs. These two years are exactly among the last three years in Figure. 5.2 when the urban area is exposing to the least heat waves among the ten years. Therefore, no matter whether there are extreme heat waves or not, it is a common phenomenon that the distribution network becomes vulnerable in CMs. The extreme weather conditions just deteriorate the severity of faults and increase the number.

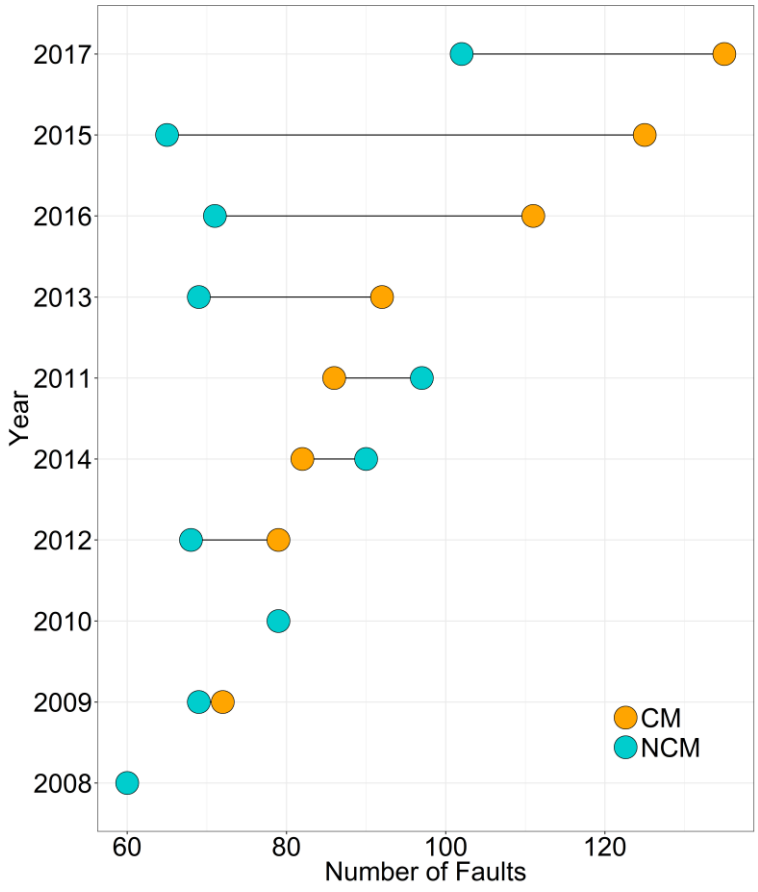


Figure 5.3 Number of outages in CM and NCM from 2008 to 2017

In this chapter, one of the novelties is to figure out the effect of extreme heat waves to the urban distribution system. As shown in Fig. 5.3, even though the sequence of years is unordered in the vertical axis, the first five years (2008-2012) are concentrated in the bottom while the latest five years are on the top. Among the first five years, the difference of interruption numbers in CMs and NCMs is relatively small. The effect of heat waves in these years is also limited as shown in Fig. 5.2. However, in the years on the top of this figure, the performance of distribution system becomes largely different. The evident gap between the number of faults in CMs and NCMs proves the vulnerability of the power grid

facing the worsening heat waves. Apart from the heat waves, the evidence of a weaker distribution system could also be seen from the figures above. As can be seen from Figure. 5.2, the severity of heat waves in 2009 is similar to that in 2016. While the difference of interruption between CMs and NCMs became much larger in 2016, which shows a lower resistance of the system to the similar weather conditions. The increasing load due to more residential air-conditioners and the aging of power equipment could both contribute to this problem. The bad consequence of heat waves could appear more frequently due to multiple reason.

Both the large number of interruptions in CMs and huge difference between CMs and NCMs in Figure 5.3 indicate that the two years of 2015 and 2017 are the year when the system suffers particular serious security issues in CMs.

5.2.3 Repetitive faults in the distribution system

With more faults happening under the extreme heat waves, a significant increase of repetitive faults has also been witnessed from the historical records. Since the repetitive fault is of low probability but will cause huge loss to the distribution network, it is attracting more attentions in the resilience problem. As one of the severe consequences in extreme heat waves, the repetitive faults will be introduced in this section.

In a typical power distribution network, the feeders are designed with the concept of single-ended power supply. To enhance the security of the system, the end of a feeder without substations is connected to the end of another feeder with an open switch as shown in Fig. 5.4. When there is no fault, the two feeders could operate independently.

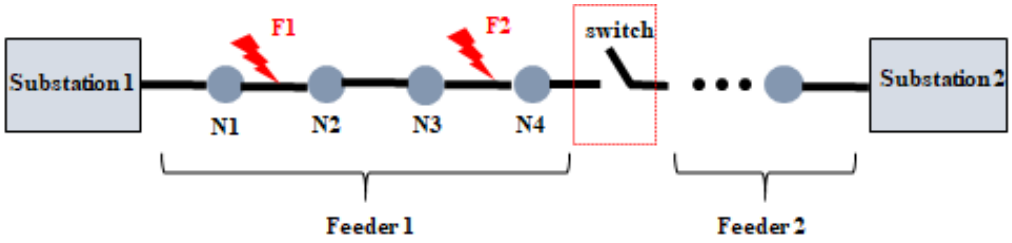


Figure 5.4 Diagram of repetitive faults in the same feeder

When a fault happens in a feeder between two nodes, the distribution system operator would cut off the power supply from Substation 1 and send a repairing team to locate the fault. For example, if only fault F1 happened as shown in Fig. 5.4, this fault could be isolated by opening the switches in nodes N1 and N2. After that, node N1 could be re-supplied once the switch inside Substation 1 is closed. All the other nodes in Feeder 1 would also be re-supplied by closing the connecting switch between the two feeders. Substation 2 will take the responsibility to supply all the nodes in the downstream of fault F1. The repairing team would take several hours to fix the problem in the power cable between N1 and N2 without losing any customers.

As discussed before, there are more interruptions expected to be witnessed in the distribution system with the attack of extreme heat waves. According to the

records from local distribution system operator, long-time blackouts are appearing more frequently. In such case, the customers have to stay without power for a relative long time with the interruption of small business and even activities in factories. The power company needs to pay a lot for the loss of its customers as well as the fine from the regulatory.

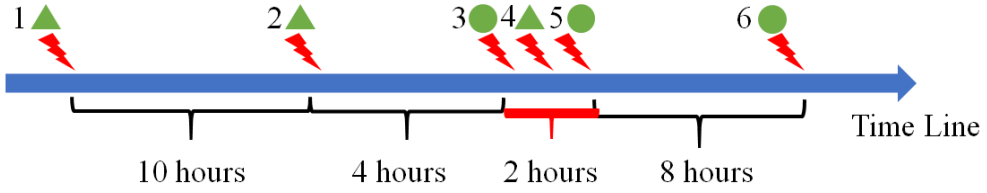


Figure 5.5 Concept of repetitive faults in time line

With intensive study of such long-time interruptions, the concept of repetitive faults is introduced, which is probably one of the principal causes. In this chapter, the repetitive fault is defined as a group of faults that happened in the same feeders within a pre-defined time threshold Δt. Correspondingly, the faults belonging to different feeders or those in the same feeder with the time difference larger than Δt are defined as non-repetitive faults. The repetitive faults can be conceptually explained in Figure 5.5, where there are 6 faults recorded within 24 hours on two different feeders labeled as green triangle and cycle, respectively.

If the time threshold for repetitive faults is set as 4 hours, among all the faults in Figure 5.5, only the third and fifth faults are belonging to one repetitive fault. Although the first and second faults happened in the same feeder, their time difference is over the pre-defined time threshold. Similarly, even though the third one happened within the time threshold after the second one, they cannot be determined as a repetitive fault due to the location of different feeders. The time threshold could be defined as the typical repairing time of a single fault. For example, if the fault F2 in Figure 5.4 happened after the restoration of the first fault F1, both of them are non-repetitive faults and there would not be any customers suffering from the long-time blackout. However, if F2 happened before the repairing of F1 is finished, the nodes N2 and N3 would be out of supply once both faults are isolated. The long-time blackout of customers connected to these two nodes could take hours until one of the faults is firstly repaired.

5.3 Cost benefit analysis

As explained in the previous section, the climate change with other factors including aging of power equipment and increasing of load is leading to a vulnerable distribution network. More interruptions as well as repetitive faults are going to be recorded under the occurrence of extreme heat waves. Urgent demand for secure power supply with a strong power grid resistant to severe weather conditions is becoming an impetus for the construction and renewal of electrical infrastructure. However, the huge investment in the fundamental infrastructure

construction is usually difficult to get enough social acceptance, which may delay or even jeopardize the projects [183].

CBA, as an essential tool to identify the effect of investments in power system, is a potential solution to make clear the contribution of projects and get the approval from regulatory. In this chapter, the performance of distribution network under extreme heat waves is studied as a resilience problem, which has low probability (2 out of 10 years) while high impact. The CBA is able to provide a robust assessment of investments about the values for society in a wide range of possible scenarios of the power system.

5.3.1 Monte Carlo simulation-based CBA

To deal with the increasingly serious problems brought by the extreme heat waves and decrease the long-time blackout to customers, the distribution project consisting of refurbishment of existing assets or addition of new equipment has been proposed. In this paper, we implement the CBA of investment on distribution network based on Monte Carlo simulation. The research will be conducted within the investment cycle, among which time the extreme heat waves may happen in multiple years under a certain probability extracted from the historical records. In Monte Carlo simulation, thousands of scenarios are generated to evaluate the benefit of investment in a pretty wide range of future cases. For each scenario, all the years in the investment cycle are divided into two parts as CMs and NCMs for a better understanding of the effect under extreme heat waves. Apparently, the cost of interruptions in CMs hit by an extreme heat wave is a typical resilience problem compared with that in CMs when no extreme heat wave happens.

For the convenience of explanation, the years with extreme heat waves are labeled as CY for critical year and those without severe weather are labeled as NCY. Therefore, the entire investment cycle could be classified into 4 periods: CYCM, CYNCM, NCYCM and NCYNCM.

According to section 5.2.3, the faults could be re-defined as repetitive faults and non-repetitive faults. Since one repetitive fault could contain several single fault records, the total number of faults in this study is actually the sum of non-repetitive and repetitive faults, which is less than the number of total records. The rate of faults could be defined with the following equation

$$\lambda = \frac{Number\ of\ faults}{(Number\ of\ months)*(length\ of\ feeders)} \tag{5.1}$$

As there are four periods during the investment cycle, the rate of faults should be calculated four times to present the different rates in four periods: λ_{ncy}^{ncm}, λ_{ncy}^{cm}, λ_{cy}^{ncm}, λ_{cy}^{cm}, for NCYNCM, NCYCM, CYNCM, CYCM, respectively. In each scenario of Monte Carlo simulation, the number of total faults is supposed to be generated according to these fault rates. Moreover, the percentages of repetitive faults in four periods could be calculated with (5.2)

$$P = \frac{Number\ of\ repetitive\ faults}{Number\ of\ total\ faults} \tag{5.2}$$

The percentage of repetitive faults is highly related to the occurrence of long-time blackout for customers. This is a factor easily affected by the extreme heat waves, which will cause large cost due to the interruption of power supply compared to the situation in normal years. Therefore, the four percentages P_{ncy}^{ncm}, $P_{ncy}^{cm}, P_{cy}^{ncm}, P_{cy}^{cm}$ are different and should be calculated independently for the four periods. An effective investment is supposed to reduce the number of both non-repetitive faults and repetitive faults. The customers and distribution system operators are going to benefit a lot from it.

5.3.2 Generation of repetitive faults and non-repetitive faults

With the fault rates λ calculated from the historical records, the expected number of faults for a specific feeder in a period before the refurbishment of power grid could be determined. Based on the expectation of faults, the number of faults in different scenarios is to be simulated with a random integer from the Poisson distribution. Poisson distribution is appropriate to describe the occurrence of events in a given period, which is suitable in our case. As a one-parameter discrete distribution, the expected number of faults is both the mean and the variance of the Poisson distribution.

For each scenario, there are supposed to be some years labeled as CYs with the others labeled as NCYs. In the NCMs of NCYs, the expected number of faults E_f in the feeder is calculated with (5.3).

$$E_f = (number\ of\ months) * \lambda_{ncy}^{ncm} * (length\ of\ feeder) \tag{5.3}$$

The number of faults in that feeder during NCYNCM of one scenario is then calculated from the Poisson distribution with E_f as the parameter. Similarly, the number of faults in NCYCM can be calculated with the fault rate λ_{ncy}^{cm}. By summing up the number of faults in NCYNCM and NCYCM, the total number of faults in each non-critical year of the scenario is obtained. These procedures are also applicable to the number of faults in critical years of the scenario.

As a distinctive feature due to the extreme heat waves, the percentage of repetitive faults before the refurbishment in different periods is calculated independently. Once the total number of faults are obtained, the number of non-repetitive faults and repetitive faults is going to be determined with the percentages calculated before. However, since the number of faults is an integer close to 0 for a typical feeder, it is impractical to directly multiply the percentage of repetitive faults with the total number, in which case almost all the scenarios of that feeder have no repetitive faults.

In this section, we will propose an algorithm to find out the years with at least one repetitive fault instead of focusing on the number of faults in a specific scenario. For example, when calculating the number of repetitive faults in NCYNCM, a group of cases are randomly selected from all the non-critical years with at least one fault according to the repetitive fault percentage P_{ncy}^{ncm}. Then, one

of the faults in each of selected year is defined as a repetitive fault. After that, the same procedure is implemented on the non-critical years with at least two faults among all the scenarios. Another fault of the selected years could be determined as a repetitive fault. These procedures would continue until the highest number of faults in the non-critical years is reached. Finally, the number of non-repetitive faults in each non-critical year is achieved by subtracting the repetitive faults from the total number.

According to our expectation, the investment will help to replace the aging equipment and renewable the essential components of the feeder, which would help the system to suffer a smaller number of interruptions. Therefore, a reduction factor should be defined to reflect the contribution of the investment. In our study, the reduction factor in CYCM is a little less that the other periods under the assumption that there are still more interruptions with the extreme heat waves even after the investment.

For a better understanding of the effect about the investment, the reduced number of faults will not be generated randomly from a Poisson distribution based on a reduced E_f. Instead, the reduction of faults could only happen in the case where there are already faults in the pre-investment scenarios. For example, a random portion of the years with at least one fault in all pre-investment scenarios are selected according to the reduction factor. One fault will be subtracted from the original number of faults in each selected case. Then, the same procedure will be applied to those years with at least two faults before the investment. This method is similar to the algorithm for determining the repetitive faults with the given percentage. It is worth to point out that the non-repetitive faults and repetitive faults after investment should be calculated independently.

Although the purpose of investment is to enhance the resilience of the distribution network, it is still possible to have more faults afterwards due to the increase of power cables. The factor LR defined in (5.4) is used to demonstrate this problem

$$LR = \frac{L_{pre}}{L_{pt}*(1-Reduction\ Factor)} \tag{5.4}$$

where L_{pre} refers to the length of cables before investment and L_{pt} refers to that after investment. If the value of LR is less than 1, the number of faults after investment is supposed to be increased. In that case, the expected number of faults E_f^{pt} could be calculated by (5.5).

$$E_f^{pt} = (number\ of\ months) * \lambda_{ncy}^{ncm} * L_{pt} * (1 - Reduction\ factor) \tag{5.5}$$

Then, the number of faults in post-investment cases is firstly generated based on the Poisson distribution. After that, the number of faults in pre-investment cases are simulated with the same algorithms introduced before. The pre-investment cases are generated by subtracting the number of faults based on the percentage ($1 - LR$).

5.3.3 Cost for non-repetitive faults

After the generation of faults in a wide range of scenarios, the benefits of investment could be calculated with cost due to different number of faults pre and post the investment. In general, there are two parts of the cost when an interruption happens: the cost for fixing and the cost for interruption of service to customers. When a non-repetitive fault happens in a feeder, the switch inside the substation would cut off the power supply of that feeder immediately. The distribution system operator would send a repairing team to confirm the location of fault from the beginning of the feeder, during which time all the customers are out of service. Once the fault is isolated, all the customers in that feeder will get power restored. As to the fault, it may take several hours for repairing without losing any customers. For a reduced number of faults, the benefits could be obtained because of the less budget for the interruptions. In this paper, the benefit for locating and repairing the non-repetitive faults after investment is defined as B_f^{NRF} with (5.6):

$$B_f^{NRF} = \left(N_{pre}^{NRF} - N_{pt}^{NRF}\right) * \left(C_{fl} * T_{fl} + C_{fr} * T_{fr}\right) \quad (5.6)$$

where N_{pre}^{NRF} and N_{pt}^{NRF} are the number of non-repetitive faults before and after the investment, respectively; C_{fl} and C_{fr} are the costs for locating and repairing the fault per hour, respectively; T_{fl} and T_{fr} are the time (in hours) for locating and repairing the fault, respectively.

As to the customers, the cost for the interruption of service is different according to their types. Usually, the cost of industrial customers is much more expensive than that of residential ones. The benefit of investment regarding to the customers can be calculated as B_c^{NRF} in (5.7)

$$B_c^{NRF} = \left(N_{pre}^{NRF} - N_{pt}^{NRF}\right) * \left(P_{ind} * T_{avg} * C_{ind} + P_{res} * T_{avg} * C_{res}\right) \quad (5.7)$$

where C_{ind} and C_{res} are the costs of interruptions for industrial and residential customers per hour per kWh, respectively; P_{ind} and P_{res} are the power of industrial and residential customers in the feeder, respectively; T_{avg} is the average outage time for all the customers along the feeder.

5.3.4 Cost for repetitive faults

As mentioned previously, when the second fault happens in the same feeder within a certain time threshold after the first one, both of the two faults are contained in one repetitive fault. Therefore, for each repetitive fault generated by the method introduced in section 5.3.2, a pair of faults have to be determined in both time and location. Since the repetitive faults will cause the customers for long-time blackout, at least one node is supposed to be in between the two single faults.

In the simulation, the generation of pair of faults inside the same repetitive fault follows the principle of randomness. Once the feeder is determined, the first

fault could happen in any part of the feeder. For a better explanation, the part of feeder between two adjacent nodes is defined as one segment. Since two single faults inside one repetitive fault could not happen in the same segment, the second fault is supposed to be located in any other segments. In the simulation, all the possible combinations of the pair of segments are used for random selection to locate the repetitive fault. For example, if there is a repetitive fault happens in the feeder in one of the scenarios, the locations of the two single faults are randomly chosen from all the possible combinations of segments, which assures one or more nodes under the threat for long-time outage. If there are more than one repetitive fault in the same scenario, another pair of segments are chosen based on the same rules.

As discussed in section 5.3.2, the number of faults after investment is generated based on a reduction factor compared with the faults before investment. In order to better analyze the cost-benefit of investment in distribution network, the pair of segments after the investment should be one of those in the pre-investment scenarios. Because if the pair of segments after investment is also chosen randomly, it may contain more nodes between the two single faults, which would bring confusion to the CBA due to more losses even with less repetitive faults.

Apart from the location, the time difference between two single faults inside one repetitive fault also plays an important role in the cost of interruptions. For example, if the second single fault happens closely to the first one, which is actually common in the historical records, the customers between the two faults would have no access to the electricity for a long time until one of the faults is fixed and the other one is isolated. However, if the second single fault happens near the end of repairing for the first one, the customers would suffer a blackout for less time since all the customers already get power restored once the first one is successfully isolated. The long-time blackout only happens when both of the single faults are isolated.

When calculating the benefit of investment regarding to the interruptions due to repetitive faults, the previous analysis for non-repetitive faults is not applicable. Because the number of customers and the time for blackout are different in the scenarios before and after investment. Therefore, the cost of interruption in different scenarios has to be calculated independently. With a long-time blackout for customers, the direct cost of a distribution system operator does not only include the fault location and repairing, but also the fees for a backup generator to support the customers in between the two single faults. This cost can be calculated as C_f^{RF} in (5.8).

$$C_f^{RF} = 2 * \left(C_{fl} * T_{fl} + C_{fr} * T_{fr}\right) + C^G \tag{5.8}$$

where C^G is the average cost of backup generator every time it is used in distribution network.

As for the cost of customers, it should be considered in both locations and time. The part of the cable between two single faults is the fault region, where the

customers suffer an extra long-time blackout except the time for fault location. This duration could be calculated from the beginning of the second fault until the restoration of the first one, denoted as T_{btw} in this paper. Since the customers in the upstream above the fault region is directly connected to the substation, once the first fault is isolated, they will get the power restored without any disturbance from the second single fault. However, the customers in the downstream of the fault region have to suffer the interruption due to fault location for two times when a repetitive fault happens.

$$C_c^{RF} = \left(P_{ind}^{UP} * C_{ind} + P_{res}^{UP} * C_{res}\right) * T_{avg} + 2 * (P_{ind}^{DN} * C_{ind} + P_{res}^{DN} * C_{res}) * T_{avg} + (P_{ind}^{RE} * C_{ind} + P_{res}^{RE} * C_{res}) * T_{btw} \quad (5.9)$$

where P_{ind}^{UP}, P_{ind}^{DN} and P_{ind}^{RE} are the total contract power of industrial customers in the upstream, downstream and fault region of the feeder, respectively; P_{res}^{UP}, P_{res}^{DN} and P_{res}^{RE} are the total contract power of residential customers in the upstream, downstream and fault region of the feeder, respectively.

Then these two costs for repetitive faults could be calculated in both pre- and post-investment scenarios, denoted with a subscript *pr* and *pt*, respectively in this paper. The benefit of investment with regarding to the cost of repairing and cost of customers in repetitive faults are calculated with the equations below

$$B_f^{RF} = C_{f,pr}^{RF} - C_{f,pt}^{RF} \quad (5.10)$$

$$B_c^{RF} = C_{c,pr}^{RF} - C_{c,pt}^{RF} \quad (5.11)$$

Finally, the benefits of investment for a specific scenario can be calculated with the Eq. (5.6), (5.7), (5.10) and (5.11).

5.4 Case study application

In this section, the proposed CBA approach is applied to a portion of the distribution network in Turin. As introduced in section 5.2, during the last 10 years, the urban area was hit by the extreme heat waves more frequently, which has become a major threat to the security of the local power grid. In this paper, the fault rates and percentage of repetitive faults in the four different periods are summarized from the historical records. All the scenarios in the Monte Carlo simulation are established based on these parameters.

In the CBA analysis, the investment cycle for our case is set as 25 years with an interest rate of 0.04. As explained in section 5.2, the probability for a critical year with extreme heat waves in Turin is 0.2 according to the 10 years' weather records. During the non-critical years and NCMs in critical years, the distribution network is under normal situation. The typical repairing time in normal situation is 6 hours while in CYCM that value is around 7 hours. The time for the repairing team to locate the fault is set as 1 hour for all the four periods in the simulation. Since the customers will get power restored from the beginning node along the feeder, most of the customers would get power restored less than the entire time for fault location. The average interruption duration of the customers in the feeder

is set as 40 minutes. According to the regulations, the cost of an industrial customer during interruption is 54 Euro/kWh, which is 4.5 times for a residential customer.

One of the scenarios about the faults happening in a portion of distribution network is shown in Figure 5.6, which is connected to the substation "S. Rita" in Turin. There are two repetitive faults (in red and green) and one non-repetitive fault (in blue) in a given period. Both industrial and residential customers could be connected to the same node.

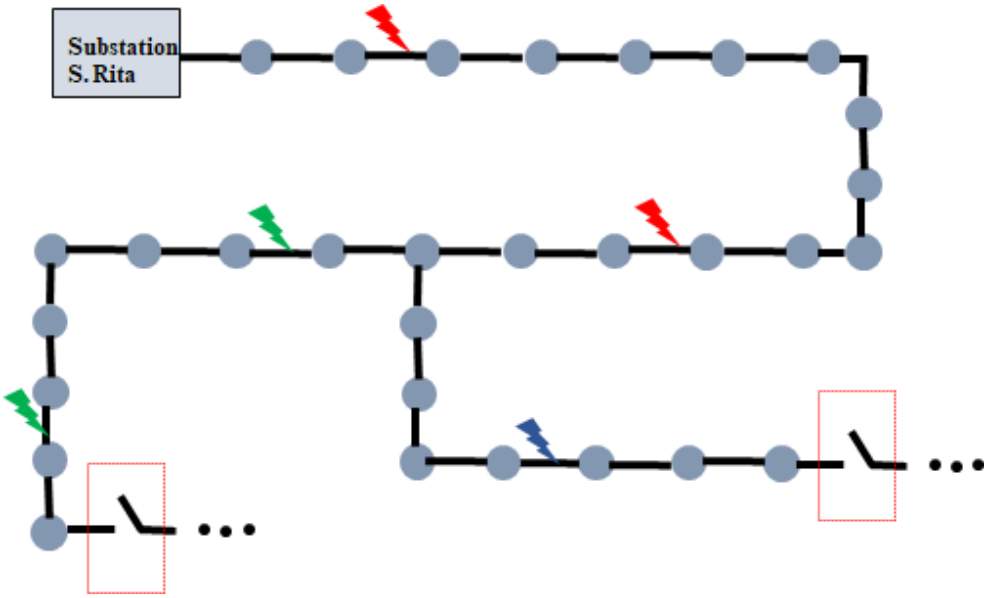


Figure 5.6 A portion of distribution network

In the Monte Carlo simulation, 100 000 scenarios are generated through the investment cycle of 25 years. By sorting these scenarios according to benefits calculated with the model in section 5.3, a curve is shown in Figure 5.7.

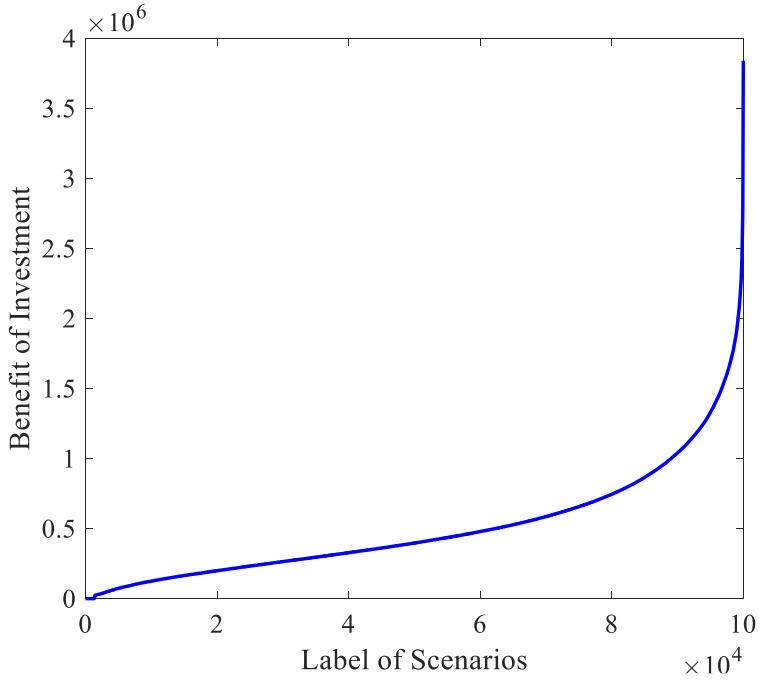


Figure 5.7 Benefit of investment in different scenarios

As can be seen from the figure above, all the benefits are positive after the investment in the portion of the network in Figure 5.6. The benefit of investment could achieve over 1.5 million Euros in the top 5% scenarios. The benefits purely in CYCMs of the scenarios are shown in Figure 5.8.

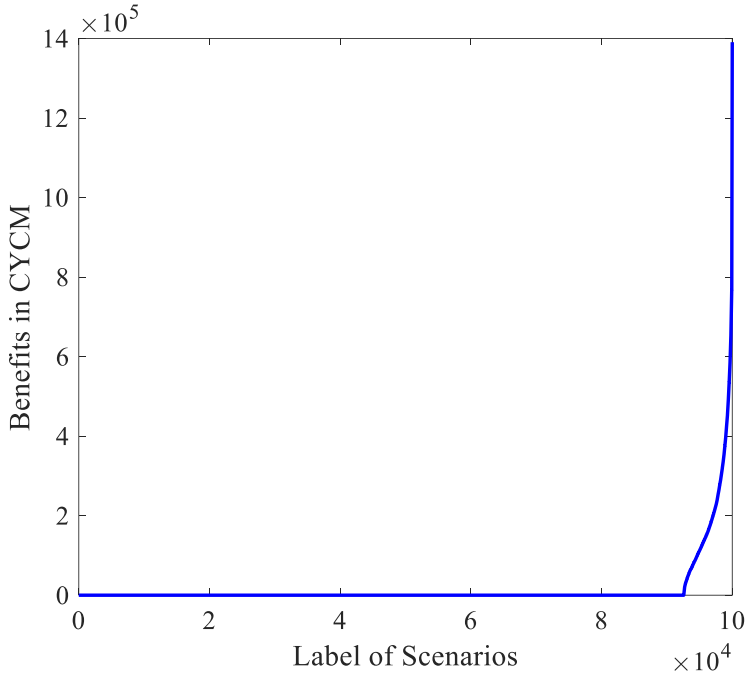


Figure 5.8 Benefit of investment only in CYCM

Since the resilience problem only happens during CYCM, Figure 5.8 is actually the benefit of improvement of the resilience in distribution network.

5.4 Summary

This chapter has proposed a Monte Carlo simulation-based model to analyze the cost-benefit of investment in the distribution network. With the increasingly serious problem of extreme heat waves in the urban area of Italy, the refurbishment of the existing power grid is becoming an urgent problem. However, the high-impact low-probability events are difficult to be modeled when evaluating the effects of the investment. The model introduced in this chapter solved this problem by calculating the benefits in a pretty wide range of scenarios. In each scenario, the cost of faults is detailed modeled with two parts: the fees for locating and fixing the fault as well as the loss of customers. The benefits of the investment are calculated by comparing the same scenario before and after the investment with a reduction factor in the faults rates.

The MT15 defined in section 5.2 is a practical indicator to evaluate the severity of weather conditions with regarding to the heat waves. Based on the weather records in 10 years, the extreme heat wave in this paper is identified with the area of MT15 over 400. Since the heat waves only happen in critical months, every year is divided into two parts for a better modeling. During the critical months suffering from extreme heat waves, the number of faults increases

distinctively with a higher percentage of repetitive faults in the distribution network. The proposed method paid a lot of efforts for the generation of these faults in four periods. The difference between the cost before and after investment is the benefit of investment.

The proposed approach has been implemented in a portion of the distribution network. All the benefits from Monte Carlo scenarios are non-negative, which is in accordance to the purpose of investment. Although the number of scenarios hit by the extreme heat waves is limited, the benefits in those scenarios are significant. With more frequent heat waves and aging assets in the distribution network, the benefits may become more obvious in the future.

Chapter 6

Conclusion and future work

This thesis provides a comprehensive overview of the applications to the data analytics methods in the power distribution system. With the rapid development of the digitalization of the electrical grid, analytical methods based on data analysis and machine learning are gradually being applied to the power system. The data analysis methods presented in this thesis cover many areas of the grid, including the forecasting of renewable energy generation, the prediction of outages in distribution network, the consumption habits of electricity consumers, as well as the cost benefit analysis of investment based on repetitive faults under heat waves. The data sources involved in this thesis are also complex, involving fault records, feeder structure information, meteorological information on different time scales, electricity consumption data of household users, etc. Through the work of feature engineering such as data integration and cleaning, the hidden patterns in the raw data are tentatively discovered and used as the research basis for the thesis topic.

In the future work, more granular data can be considered for modeling. For example, the feeder condition monitoring data before and after a failure could be combined with weather information for prediction from both internal and external causes. The complex connections of the lines in power grid should also be considered while identifying the typical feeder characteristics.

References

[1] Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A survey, Computer Science Review, 2015, Volume 17, pp. 70–81

[2] Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, and Minho JoNex-Generation Big Data Analytics: State of the Art, Challenges and Future Research Topics. IEEE Transactions on Industrial Informatics, volume 13, no. 4, pp. 1891-1899.

[3] Wendy Arianne Günther, Mohammad H. Rezazade Mehrizi, Marleen Huysman, Frans Feldberg: Debating Big Data: A Literature Review on Realizing Value from Big Data. Journal of Strategic Information Systems, volume 26, pp. 191-209.

[4] Yu Changhui, Pan Heping. Business intelligence and its key technology. Application Research of Computers, 2002(9): 14-16.

[5] Ronay Ak, Olga Fink, and Enrico Zio. Two Machine Learning Approaches for Short-Term Wind Speed Time-Series Prediction. IEEE Transactions on Neural Networks and Learning Systems (Volume: 27, Issue: 8, Aug. 2016): 1734 – 1747.

[6] Wenbin Wu and Mugen Peng. A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting. IEEE Internet of Things Journal (Volume: 4, Issue: 4, Aug. 2017): 2327-4662.

[7] Ye Ren, Ponnuthurai Nagaratnam Suganthan, and Narasimalu Srikanth. A Novel Empirical Mode Decomposition with Support Vector Regression for Wind Speed Forecasting. IEEE Transactions on Neural Networks and Learning Systems (Volume: 27, Issue: 8, Aug. 2016): 1793 – 1798.

[8] Chunming Tu, Xi He, Zhikang Shuai and Fei Jiang. Big Data Issues in Smart Grid – A Review. Renewable and Sustainable Energy Reviews, volume 79, pp. 1099-1107.

[9] P. Zikopoulos, C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Education, 2011.

[10] Stephen Kaisler, Frank Amnour, J. Alberto," Big Data: Issues and Challenges Moving Forward",46th Hawaii International Conference on System Science, IEEE,2012

[11] SmartGrids European Tech. Platform, Strategic Deployment Document for Europe's Electricity Networks of the Future 6 (2010) [hereinafter E.U. SmartGrids SDD].

[12] Zhen Zhang. Smart Grid in America and Europe: Similar Desires, Different Approaches. Public Utilities Fortnightly, Vol. 149, No. 1, January 1, 2011

[13] Executive Office of the President. Economic Benefits Of Increasing Electric Grid Resilience To Weather Outages. USA. August 2013

[14] CEN-CENELEC-ETSI Smart Grid Working Group Reference Architecture, “Reference Architecture for the Smart Grid,” Tech. Rep., 2012.
[15] Ting Zhu, Sheng Xiao, Qingquan Zhang, Yu Gu, Ping Yi, and Yanhua Li, “Emergent Technologies in Big Data Sensing: A Survey”, International Journal of Distributed Sensor Networks, Volume 2015, Article ID 902982
[16] Seref SAGIROGLU, Ramazan TERZI, Yavuz CANBAY, Ilhami COLAK. Big Data Issues in Smart Grid Systems. 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), Birmingham, UK, 20-23 Nov. 2016.
[17] (June 23, 2016) Big Data analytics and energy consumption. Available: http://www.lavastorm.com/blog/2012/04/09/big-data-analytics-and-energy-consumption/
[18] Liu Keyan, Sheng Wanxin, Zhang Dongxia, et al. Big Data Application Requirements and Scenario Analysis in Smart Distribution Network, Proceedings of the CSEE, Vol. 35, No. 2, pp. 287-293, 2015
[19] Zhao Teng, Zhang Yan and Zhang Dongxia. Application Technology of Big Data in Smart Distribution Grid and Its Prospect Analysis. Power System Technology, vol. 38, no. 12, pp. 3305-3312, 2014
[20] Qang Jiye, Ji Zhixiang, Shi Mengjie, et al. Scenario Analysis and Application Research on Big Data in Smart Power Distribution and Consumption Systems. Proceedings of the CSEE, vol. 35, no. 8, pp. 1829-1836, 2015
[21] D. Baimel, S. Tapuchi, N. Baimel. Smart grid communication technologies- overview, research challenges and opportunities. International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 22-24 June 2016.
[22] Alireza Bahmanyar, Sadegh Jamali, Abouzar Estebsari, Enrico Pons, Ettore Bompard, Edoardo Patti, Andrea Acquaviva. Emerging Smart Meters in Electrical Distribution Systems: Opportunities and Challenges. 24th Iranian Conference on Electrical Engineering (ICEE), 10-12 May 2016, Shiraz, Iran.
[23] Cheng Fan, Fu Xiao, Zhengdao Li, Jiayuan Wang. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings, volume. 159, pp. 296-308, 2018.
[24] Ying Cheng, Ken Chen, Hemeng Sun, Yongping Zhang, Fei Tao. Data and Knowledge Mining with Big Data towards Smart Production. Journal of Industrial Information Integration, volume 9, pp. 1-13, 2018.
[25] Anandarup Roya, Rafael M.O. Cruz a, Robert Sabourina, George D.C. Cavalcanti. A study on combining dynamic selection and data preprocessing for imbalance learning. Neurocomputing, volume 286, pp. 179-192, 2018.
[26] Ee-Peng Lim, Jaideep Srivastava, Satya Prabhakar, James Richardson, Entity identification in database integration, In Information Sciences, Volume 89, Issues 1–2, 1996, Pages 1-38, ISSN 0020-0255.
[27] Di Zhua, Theodoros Lappasa, Juheng Zhang. Unsupervised Tip-mining from Customer Reviews. Decision Support Systems, volume 107, pp. 116-124, 2018.

[28] Joseph Siryani, Bereket Tanju, and Timothy J. Eveleighi. A Machine Learning Decision-Support System Improves the Internet of Things' Smart Meter Operations. Accident Analysis and Prediction, volume 4, pp. 1056-1066, 2017.
[29] Debi Prasad Mishra, Subhransu Ranjan Samantaray and Geza Joos. A combined wavelet and data-mining based intelligent protection scheme for microgrid. IEEE Transactions on Smart Grid 2016; 7(5): 2295 – 2304.
[30] Susmita Kar, S. R. Samantaray, and M. Dadash Zadeh. Data-mining Model based Intelligent Differential Microgrid Protection Scheme. IEEE Systems Journal (Volume: 11, Issue: 2, June 2017): 1161 – 1169.
[31] Farid Hashemi, Mohammad Mohammadi and Amin Kargarian. Islanding Detection Method for Microgrid based on Extracted Features from Differential Transient Rate of Change of Frequency. IET Generation, Transmission & Distribution (Volume: 11, Issue: 4, 3 9 2017): 891 – 904.
[32] Mollah Rezaul Alam, Kashem M. Muttaqi, Abdesselam Bouzerdoum. Evaluating the Effectiveness of A Machine Learning Approach based on Response Time and Reliability for Islanding Detection of Distributed Generation. IET Renewable Power Generation (Volume: 11, Issue: 11, 9 13 2017).
[33] Konstantin Bauman, Alexander Tuzhilin, and Ryan Zaczynski. Using Social Sensors for Detecting Emergency Events: A Case of Power Outages in the Electrical Utility Industry. ACM Transactions on Management Information Systems (Volume: 8, Issue: 2-3, 2017)
[34] Haifeng Sun, Zhaoyu Wang, Jianhui Wang, Zhen Huang, NichelleLe Carrington, and Jianxin Liao. Data-Driven Power Outage Detection by Social Sensors. IEEE Transactions on Smart Grid (Volume: 7, Issue: 5, 2016): 2516-2524.
[35] Xiaoyu Wang, Stephen McArthur, Scott Strachan, John Kirkwood and Bruce Paisley. A data analytic approach fault diagnosis and prognosis for distribution automation. IEEE Transactions on Smart Grid 2017.
[36] Enrico De Santis, Lorenzo Livi, Alireza Sadeghian, Antonello Rizzi, Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification, In Neurocomputing, Volume 170, 2015, Pages 368-383, ISSN 0925-2312.
[37] Enrico De Santis, Antonello Rizzi, Alireza Sadeghian. A Learning Intelligent System for Classification and Characterization of Localized Faults in Smart Grids. 2017 IEEE Congress on Evolutionary Computation (CEC), San Sebastian, Spain, July 2017.
[38] Yixin Cai and Mo-Yuen Chow. Exploratory Analysis of Massive Data for Distribution Fault Diagnosis in Smart Grids. IEEE Conference on Power & Energy Society General Meeting, July 2009.
[39] Jian Wang, Xiaofu Xiong, Ning Zhou, Zhe Li, Wei Wang. Early Warning Method for Transmission Line Galloping based on SVM and AdaBoost bi-level Classifiers. IET Generation, Transmission & Distribution (Volume: 10, Issue: 14, 11 4 2016): 3499 – 3507.
[40] Yuchen Zhang, Yan Xu, Zhao Yang Dong, Zhao Xu, and Kit Po Wong. Intelligent Early Warning of Power System Dynamic Insecurity Risk_Toward

Optimal Accuracy-Earliness Tradeoff. IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 5, Oct. 2017): 2544 – 2554.
[41] Qiushi Cui, Khalil El-Arroudi and G´eza Jo´os. An Effective Feature Extraction Method in Pattern Recognition Based High Impedance Fault Detection. 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP), San Antonio, TX, USA, September 2017.
[42] Huaiguang Jiang, Xiaoxiao Dai, David Wenzhong Gao, Jun Jason Zhang, Yingchen Zhang, and Eduard Muljadi. Spatial-Temporal Synchrophasor Data Characterization and Analytics in Smart Grid Fault Detection Identification and Impact Casual Analysis. IEEE Transactions on Smart Grid (Volume: 7, Issue: 5, Sept. 2016): 2525 – 2536.
[43] Chengxi Liu, Kai Sun, Zakir Hussain Rather, Zhe Chen, ClausLethBak, Paul Thøgersen and Per Lund. A Systematic Approach for Dynamic Security Assessment. IEEE Transactions on Power Systems (Volume: 29, Issue: 2, March 2014): 717-730.
[44] Miao He, Junshan Zhang and Vijay Vittal. Robust online dynamic security Assesment Using Adaptive Ensemble Decision-Tree Learning. IEEE Transactions on Power Systems (Volume: 28, Issue: 4, Nov. 2013): 4089 – 4098.
[45] Chuyao He and Lin Guan, Weike Mo. A Method for Transient Stability Assessment based on Pattern Recognition. International Conference on Smart Grid and Clean Energy Technologies (ICSGCE), Oct. 2016.
[46] Madhumita Parate, Sachin Tajane and Bhagyashri Indi. Assessment of System Vulnerability for Smart Grid Applications. IEEE International Conference on Engineering and Technology (ICETECH), March 2016.
[47] Teodora Dimitrovska, Urban Rudež, Rafael Mihalič. Fast Contingency Screening Based on Data Mining. IEEE EUROCON International Conference on Smart Technologies, Ohrid, Macedonia, 2017.
[48] Chowdhury Andalib-Bin-Karim, Xiaodong Liang, Nahidul Khan and Huaguang Zhang. Determine Q-V Characteristics of Grid-connected Wind Farms for Voltage Control using a Data-driven Analytics Approach. IEEE Transactions on Industry Applications (Volume: 53, Issue: 5, Sept.-Oct. 2017).
[49] Aleena Swetapadma and Anamika Yadav. Data-mining-based fault during power swing identification in power transmission system. IET Science, Measurement & Technology (Volume: 10, Issue: 2, 3 2016): 130 – 139
[50] Manas Kumar Jena and Subhransu Ranjan Samantaray. Data-mining-based Intelligent Differential Relaying for Transmission Lines Including UPFC and Wind Farms. IEEE Transactions on Neural Networks and Learning Systems (Volume: 27, Issue: 1, Jan. 2016): 8 – 17.
[51] Panagiotis N. Papadopoulos, Tingyan Guo, and Jovica V. Milanović. Probabilistic Framework for Online Identification of Dynamic Behavior of Power Systems with Renewable Generation. IEEE Transactions on Power Systems (Volume: PP, Issue: 99):
[52] Lipeng Zhu, Chao Lu, Zhao Yang Dong, and Chao Hong. Imbalance Learning Machine-based Power System Short-term Voltage Stability Assessment.

IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 5, Oct. 2017): 2533 - 2543.
[53] LIU Wei, ZHANG Dongxia, WANG Xinying, LIU Daowei，WU Qian. Power System Transient Stability Analysis Based on Random Matrix Theory. Proceedings of the CSEE, Vol.36 No.18, pp: 4854-4863, Sep. 20, 2016.
[54] XUXinyi, HEXing, AIQian, QIUCaiming. A Correlation Analysis Method for Operation Status of Distribution Network Based on Random Matrix Theory. Power System Technology, Vol. 40 No. 3, pp: 781-790, Mar. 2016
[55] Junbo Zhang, C. Y. Chung, Zejing Wang, and Xiangtian Zheng. Instantaneous Electromechanical Dynamics Monitoring in Smart Transmission Grid. IEEE Transactions on Industrial Informatics (Volume: 12, Issue: 2, April 2016): 844 – 852.
[56] Vuk Malbasa, Ce Zheng, Po-Chen Chen, Tomo Popovic and Mladen Kezunovic. Voltage Stability Prediction Using Active Machine Learning. IEEE Transactions on Smart Grid (Volume: 8, Issue: 6, Nov. 2017): 3117 – 3124.
[57] Junbo Zhao, Gexiang Zhang, Kaushik Das, George N. Korres, Nikolaos M. Manousakis, Avinash K. Sinha, and Zhengyou He. Power System Real-Time Monitoring by Using PMU-based Robust State Estimation Method. IEEE Transactions on Smart Grid (Volume: 7, Issue: 1, Jan. 2016): 300 – 309.
[58] Zubair Shah, Adnan Anwar, Abdun Naser Mahmood, Zahir Tari and Albert Y. Zomaya. A Spatio-temporal Data Summarization Paradigm for Real-time Operation of Smart Grid. IEEE Transactions on Big Data (Volume: PP, Issue: 99), 2017.
[59] Bo Wang, Member, Biwu Fang, Yajun Wang, Hesen Liu, and Yilu Liu. Power System Transient Stability Assessment based on Big Data and the Core Vector Machine. IEEE Transactions on Smart Grid (Volume: 7, Issue: 5, Sept. 2016): 2561 – 2570.
[60] Jiaqing Lv, Mirosław Pawlak, and U. D. Annakkage. Prediction of the Transient Stability Boundary based on Nonparameteric Additive Modeling. IEEE Transactions on Power Systems (Volume: 32, Issue: 6, Nov. 2017): 4362 – 4369.
[61] Andreas Reinhardt and Delphine Reinhardt. Detecting Anomalous Electrical Appliance Behavior based on Motif Transition Likelihood Matrices. 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, NSW, Australia, Sydney, NSW, Australia.
[62] Gehao Sheng, Huijuan Hou, Xiuchen Jiang and Yufeng Chen. A Novel Association Rule Mining Method of Big Data for Power Transformer State Parameters based on Probabilistic Graph Model. IEEE Transactions on Smart Grid (Volume: PP, Issue: 99).
[63] William Hand Allen, Ahmed Rubaai, and Ramesh Chawla. Fuzzy Neural Network-based Health Monitoring for HVAC System Variable-Air-Volume Unit. IEEE Transactions on Industry Applications (Volume: 52, Issue: 3, May-June 2016): 2513 – 2524.
[64] D. N. P. Murthy, M. Bulmer, and J. A. Eccleston, “Weibull model selection for reliability modeling,” Rel. Eng. Syst. Safety, vol. 86, no. 3, pp. 257–267, Dec. 2004.

[65] Jian Qiu, Huifang Wang, Dongyang Lin, Benteng He, Wanfang Zhao, and Wei Xu. Nonparameteric Regression-based Failure Rate Model for Electric Power Equipment Using Lifecycle Data. 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Dallas, TX, USA, May 2016.
[66] A. G. Colombo, D. Costantini, and R. J. Jaarsma, "Bayes nonparametric estimation of time-dependent failure rate," IEEE Trans. Rel., vol. 34, no. 2, pp. 109–112, Jun. 1985.
[67] Ebrahim Balouji and Ozgul Salor. Classification of Power Quality Events Using Deep Learning on Events Images. 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), April 2017.
[68] Yachen Tang, Chee-Wooi Ten, Chaoli Wang and Gordon Parker. Extraction of Energy Information from Analog Meters using Image Processing. IEEE Transactions on Smart Grid (Volume: 6, Issue: 4, July 2015).
[69] Fábbio A. S. Borges, Ricardo A. S. Fernandes, Ivan N. Silva, and Cíntia B. S. Silva. Feature Extraction and Power Quality Disturbances Classification using Smart Meters Signals. IEEE Transactions on Industrial Informatics (Volume: 12, Issue: 2, April 2016): 824 – 833.
[70] Ferhat UÇAR, Ömer Faruk ALÇİN, Beşir DANDIL and Fikret ATA. Machine Learning based Power Quality Event Classification using Wavelet_Entropy and Basic Statisticsal Features. 2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR), August 2016, Miedzyzdroje, Poland.
[71] V. Gungor et al., "Smart grid technologies communications technologies and standards," IEEE Trans. Ind. Informat., vol. 7, no. 4, pp. 529–539, Sep. 2011.
[72] D. Ghosh, T. Ghose, and D. K. Mohanta, "Communication feasibility analysis for smart grid with phasor measurement units," IEEE Trans. Ind. Informat., vol. 9, no. 3, pp 1486–1496, Aug. 2013.
[73] V. Kekatos, G. B. Giannakis, and R. Baldick, "Grid topology identification using electricity prices," in Proc. IEEE Power Energy Soc. Gen. Meeting, National Harbor, MD, USA, Jul. 2014, pp. 1–5.
[74] Mohammad Babakmehr, Marcelo G. Simões, Michael B. Wakin, and Farnaz Harirchi. Compressive Sensing-based Topology Identification for Smart Grids. IEEE Transactions on Industrial Informatics (Volume: 12, Issue: 2, April 2016): 532 – 543.
[75] Ran Li, Chenghong Gu, Furong Li, Gavin Shaddick, and Mark Dale. Development of Low Voltage Network Templates_Part I_Substation Clustering and Classification. IEEE Transactions on Power Systems (Volume: 30, Issue: 6, Nov. 2015).
[76] Ran Li, Chenghong Gu, Furong Li, Gavin Shaddick, and Mark Dale. Development of Low Voltage Network Templates—Part II_ Peak Load Estimation by Clusterwise Regression. IEEE Transactions on Power Systems (Volume: 30, Issue: 6, Nov. 2015).
[77] Ming Yang, You Lin and Xueshan Han. Probabilistic Wind Generation Forecast Based on Sparse Bayesian Classification and Dempster-Shafer Theory.

2015 IEEE Industry Applications Society Annual Meeting, Addison, TX, USA, Oct. 2015.
[78] Mahdi Khodayar, Okyay Kaynak and Mohammad E. Khodayar. Rough Deep Neural Architecture for Short-term Wind Speed Forecasting. IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 6, Dec. 2017): 2770 – 2779.
[79] TENG ZHAO, ZIQIANG ZHOU, YAN ZHANG, PING LING, AND YINGJIE TIAN. Spatio-temporal Analysis and Forecasting of Distributed PV Systems Diffusion_A Case Study of Shanghai Using A Data-driven Approach. IEEE Access (Volume: 5): 5135 – 5148, 2017.
[80] H. Nazaripouya, B. Wang, Y. Wang, P. Chu, H. R. Pota, and R. Gadh. Univariate Time Series Prediction of Solar Power Using a Hybrid Wavelet-ARMA-NARX Prediction Method. 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Dallas, TX, USA.
[81] D. W. Bunn and E. D. Farmer, ComparativeModels for Electrical Load Forecasting. New York, NY, USA: Wiley, 1985, Chichester.
[82] Ni Ding, Clémentine Benoit, Guillaume Foggia, Yvon Bésanger, Senior Member, IEEE, and Frédéric Wurt. Neural Network-Based Model Design for Short-Term Load Forecast in Distribution Systems. IEEE Transactions on Power Systems (Volume: 31, Issue: 1, Jan. 2016): 72 – 81.
[83] Dunnan Liu, Long Zeng, Canbing Li, Kunlong Ma, Yujiao Chen and Yijia Cao. IEEE Systems Journal (Volume: PP, Issue: 99).
[84] Song Li, Peng Wang, and Lalit Goel. A Novel Wavelet-Based Ensemble Method for Short-Term Load Forecasting with Hybrid Neural Networks and Feature Selection. IEEE Transactions on Power Systems (Volume: 31, Issue: 3, May 2016): 1788 – 1798.
[85] Ashfaq Ahmad, Nadeem Javaid, Mohsen Guizani, Nabil Alrajeh, and Zahoor Ali Khan. An Accurate and Fast Converging Short-Term Load Forecasting Model for Industrial Applications in a Smart Grid. IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 5, Oct. 2017): 2587 – 2596.
[86] Heng Shi, Minghao Xu and Ran Li. Deep Learning for Household Load Forecasting – A Novel Pooling Deep RNN. IEEE Transactions on Smart Grid (Volume: PP, Issue: 99)
[87] Weicong Kong, Zhao Yang Dong, Youwei Jia, David J. Hill, Yan Xu and Yuan Zhang. Short-Term Residential Load Forecasting based on LSTM Recurrent Neural Network. IEEE Transactions on Smart Grid (Volume: PP, Issue: 99).
[88] Ding Li, and Sudharman K. Jayaweera. Machine-learning Aided Optimal Customer Decision for an Interactive Smart Grid. IEEE Systems Journal (Volume: 9, Issue: 4, Dec. 2015): 1529 – 1540.
[89] A. Moreno-Munoz, F.J. Bellido-Outeirino, P. Siano, M.A. Gomez-Nieto. Mobile social media for smart grids customer engagement: Emerging trends and challenges. Renewable and Sustainable Energy Reviews, vol. 53, 2016: 1611-1616.

[90] Y. Cai, T. Huang, E. Bompard, Y. Cao and Y. Li, "Self-Sustainable Community of Electricity Prosumers in the Emerging Distribution System," in IEEE Transactions on Smart Grid, vol. 8, no. 5, pp. 2207-2216
[91] PENG Wenhao, DENG Zhe, ZHU Yanping, LU Jun. An Analytical Method for Intelligent Electricity Use Pattern with Demand Response. 2016 China International Conference on Electricity Distribution (CICED), Xi'an, China, Aug. 2016.
[92] X. Liu, S. Bolwig, and P.S. Nielsen. SmartM: A Non-intrusive Load Monitoring Platform, Proc. of the 22nd Business Information System (Workshop), pp. 424-434, 2019.
[93] Ramon Granell, Colin J. Axon, and David C. H. Wallom. Impact of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. IEEE Transactions on Power Systems (Volume: 30, Issue: 6, Nov. 2015): 3217 – 3224.
[94] Shailendra Singh and Abdulsalam Yassine. Mining Energy Consumption Behavior Patterns for Households in Smart Grid. IEEE Transactions on Emerging Topics in Computing (Volume: PP, Issue: 99).
[95] P. Gianniou, X. Liu*, A. Heller, P. S. Nielsen, and C. Rode. Clustering-based Analysis for Residential District Heating Data. Journal of Energy Conversion & Management, vol. 165, pp. 840-850, 2018.
[96] Nadia Ahmed, Marco Levorato, and G.P. Li. Residential Consumer-Centric Demand Side Management. IEEE Transactions on Smart Grid (Volume: PP, Issue: 99).
[97] Bo Peng, Can Wan, Shufeng Dong, Jin Lin, Yonghua Song, Yi Zhang, Jun Xiong. A Two-stage Pattern Recognition Method for Electric Customer Classification in Smart Grid. 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, NSW, Australia, Nov. 2016.
[98] Reem Al-Otaibi, Nanlin Jin, Member, IEEE, Tom Wilcox, and Peter Flach. Feature Construction and Calibration for Clustering Daily Load Curves from Smart-meter Data. IEEE Transactions on Industrial Informatics (Volume: 12, Issue: 2, April 2016): 645 – 654.
[99] Shiyin Zhong, and Kwa-Sur Tam. A Frequency Domain Approach to Characterize and Anlyze Load Profiles. IEEE Transactions on Power Systems (Volume: 27, Issue: 2, May 2012): 857 – 865.
[100] Ran Li, Furong Li, and Nathan D. Smith. Load Characterization and Low-order Approximation for Smart Metering Data in the Spectral Domain. IEEE Transactions on Industrial Informatics (Volume: 13, Issue: 3, June 2017): 976 - 984.
[101] Dong Zhang, Shuhui Li, Min Sun, and Zheng O'Neill. An Optimal and Learning-based Demand Response and Home Energy Management System. IEEE Transactions on Smart Grid (Volume: 7, Issue: 4, July 2016): 1790 – 1801.
[102] Anish Jindal, Amit Dua, Kuljeet Kaur, Mukesh Singh, Neeraj Kumar, and S. Mishra. Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid. IEEE Transactions on Industrial Informatics, vol. 12, no. 3, june 2016: 1005-1016.

[103] Stephen Haben, Colin Singleton, and Peter Grindrod. Analysis and Clustering of Residential Customers Energy Behavioral Demand using Smart Meter Data. IEEE Transactions on Smart Grid (Volume: 7, Issue: 1, Jan. 2016): 136 – 144.
[104] Sergio Valero Verdú, Mario Ortiz García, Carolina Senabre, Antonio Gabaldón Marín, and Francisco J. García Franco. Classification Filtering and Identification of Electrical Customer Load Patterns through the Use of Self-Organizing Maps. IEEE Transactions on Power Systems (Volume: 21, Issue: 4, Nov. 2006): 1672 – 1682.
[105] Ran Li, Furong Li and Nathan D. Smith. Multi-Resolution Load Profile Clustering for Smart Metering Data. IEEE Transactions on Power Systems (Volume: 31, Issue: 6, Nov. 2016): 4473 – 4482.
[106] Amr A. Munshi, and Yasser Abdel-Rady I. Mohamed. Extracting and Defining Flexibility of Residential Electrical Vehicle Charging Loads. IEEE Transactions on Industrial Informatics (Volume: PP, Issue: 99).
[107] Xing Tong, Chongqing Kang and Qing Xia. Smart Metering Load Data Compression based on Load Feature Identification. IEEE Transactions on Smart Grid (Volume: 7, Issue: 5, Sept. 2016): 2414 – 2422.
[108] Yi Wang, Qixin Chen, Chongqing Kang, Qing Xia, and Min Luo. Sparse and Redundant Representation-based Smart Meter Data Compression and Pattern Extraction. IEEE Transactions on Power Systems (Volume: 32, Issue: 3, May 2017): 2142 – 2151.
[109] Jian Liang, Simon K. K. Ng, Gail Kendall, and John W. M. Cheng. Load Signature Study-Part I: Basic Concept Structure and Methodology. IEEE Transactions on Power Delivery (Volume: 25, Issue: 2, April 2010): 551 – 560.
[110] Jian Liang, Simon K. K. Ng, Gail Kendall, and John W. M. Cheng. Load Signature Study-Part II: Disaggregation Framework Simulation and Applications. IEEE Transactions on Power Delivery (Volume: 25, Issue: 2, April 2010): 561 – 569.
[111] Jessie M. Gillis, Sami M. Alshareef, and Walid G. Morsi. Nonintrusive Load Monitoring Using Wavelet Design and Machine Learning. IEEE Transactions on Smart Grid (Volume: 7, Issue: 1, Jan. 2016): 320 – 328.
[112] F. Sultanem. Using appliance signatures for monitoring residential loads at meter panel level. IEEE Transaction on Power Delivery, 6(4), 1991.
[113] M. Berges, E. Goldman, H. S. Matthews, and L Soibelman. Learning systems for electric comsumption of buildings. In ASCI International Workshop on Computing in Civil Engineering, 2009.
[114] W. Lee, G. Fung, H. Lam, F. Chan, and M. Lucente. Exploration on load signatures. International Conference on Electrical Engineering (ICEE), 2004.
[115] Weicong Kong, Zhao Yang Dong, Jin Ma, David J. Hill, Junhua Zhao, and Fengji Luo. An Extensible Approach for Non-Intrusive Load Disaggregation with Smart Meter Data. IEEE Transactions on Smart Grid (Volume: PP, Issue: 99).
[116] Nilson Henao, Kodjo Agbossou, Sousso Kelouwani, Yves Dubé, and Michaël Fournier. Approach in Nonintrusive Type I Load Monitoring Using

Subtractive Clustering. IEEE Transactions on Smart Grid (Volume: 8, Issue: 2, March 2017): 812 – 821.
[117] Shikha Singh and Angshul Majumdar. Deep Sparse Coding for Non-intrusive Load Monitoring. IEEE Transactions on Smart Grid (Volume: PP, Issue: 99).
[118] J. Zico Kolter, Siddarth Batra and Andrew Y. Ng. Energy Disaggregation via Discriminative Sparse Coding. NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, Pages 1153-1161, Vancouver, British Columbia, Canada — December 06 - 09, 2010.
[119] Jessie M. Gillis, and Walid G. Morsi. Non-intrusive Load Monitoring Using Semi-supervised Machine Learning and Wavelet Design. IEEE Transactions on Smart Grid (Volume: 8, Issue: 6, Nov. 2017): 2648 – 2655.
[120] Paria Jokar, Nasim Arianpoo, and Victor C. M. Leung. Electricity Theft Detection in AMI Using Customers' Consumption Patterns. IEEE Transactions on Smart Grid, vol. 7, no. 1, january 2016: 216-226.
[121] Tung-Sheng Zhan, Shi-Jaw Chen, Chih-Cheng Kao, Chao-Lin Kuo, Jian-Liung Chen, Chia-Hung Lin. Non-technical loss and power blackout detection under advanced metering infrastructure using a cooperative game based inference mechanism. IET Gener. Transm. Distrib., 2016, Vol. 10, Iss. 4, pp. 873–882.
[122] Non-Cooperative Game Model Applied to an Advanced Metering Infrastructure for Non-Technical Loss Screening in Micro-Distribution Systems. IEEE Transactions on Smart Grid, vol. 5, no. 5, september 2014: 2468-2469.
[123] J.I. Guerrero, Iñigo Monedero, Félix Biscarri, Jesús Biscarri, Rocío Millán, Carlos León. Non-Technical Losses Reduction by Improving the Inspections Accuracy in a Power Utility. IEEE Transactions on Power Systems. (Early Access)
[124] PR Newswire. (2014). "World Loses $89.3 Billion to Electricity Theft Annually, $58.7 Billion in Emerging Markets." [Online]. Available: http://www.prnewswire.com/news-releases/world-loses-893-billion-to-electricity-theft-annually-587-billion-in-emerging-markets-300006515.html, Accessed on: Jul. 2015.
[125] Marcelo Zanetti, Edgard Jamhour, Marcelo Pellenz, Manoel Penna, Voldi Zambenedetti, Ivan Chueiri. A Tunable Fraud Detection System for Advanced Metering Infrastructure using Short-Lived Patterns. IEEE Transactions on Smart Grid. (Early Access)
[126] Y. Wang, J. Zhou, and Z. Li, et al, "Discriminant-Analysis-Based Single-Phase Earth Fault Protection Using Improved PCA in Distribution Systems," IEEE Trans on Power Delivery, vol. 30, no. 4, pp. 1974-1982, 2015.
[127] S. Sachan, C. Zhou and R. Wen, et al, "Multiple Correspondence Analysis to Study Failures in a Diverse Population of a Cable," IEEE Trans on Power Delivery, vol. 32, no. 4, pp. 1696-1704, 2016.
[128] Rigoni V, Ochoa L F and Chicco G, et al. "Representative Residential LV Feeders: A Case Study for the North West of England," IEEE Trans. Power Syst., vol. 31, no. 1, pp.348-360, 2016.

[129] A. K. Dashtaki, M. R. Haghifam, "A New Loss Estimation Method in Limited Data Electric Distribution Networks," IEEE Trans on Power Delivery, vol. 28, no. 4, pp. 2194-2200, 2013.
[130] K. P. Schneider, Y. Chen, D. Engle and D. Chassin, "A Taxonomy of North American Radial Distribution Feeders," in IEEE Power & Energy Society General Meeting, pp. 1-6, 2009.
[131] Y. L. Li amd P. J. Wolfs, "A Statistical Study on Topological Features on High Voltage Distribution Networks in Western Australia," Universities Power Engineering Conference IEEE, pp.1-6, 2010.
[132] Y. L. Li and P. J. Wolfs, "Taxonomic Description for Western Australian Distribution Medium-voltage and Low-voltage Feeders," IET Gener. Transm. Distrib., vol. 8, no. 1, pp. 104-113, 2014.
[133] R. H. Broderick, J. R. Williams, "Clustering Methodology for Classifying Distribution Feeders," Photovoltaic Specialists Conference IEEE, pp. 1706-1710, 2013.
[134] H. H. Hsu and C. W. Hsieh, "Feature Selection via Correlation Coefficient Clustering," Journal of Software, vol. 5, no. 12, 2010.
[135] Y. Chen, J. Y. Wen and S. J. Cheng, "Probabilistic load flow method based on Nataf transformation and Latin hypercube sampling," IEEE Trans. Sustain. Energy, vol. 4, no. 2, pp. 294-301, 2013.
[136] P. N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining," Data Analysis in the Cloud, vol. 26, no. 25, pp. 1-25, 2016.
[137] R.J.F. Ypma and W.M.V Ballegooijen, "A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes," Epidemiology, vol. 24, no. 3, pp. 395-400, 2013.
[138] N. S. Kumar, K. N. Rao and A. Govardhan, et al. "Undersampled K-means Approach for Handling Imbalanced Distributed Data," Progress in Artificial Intelligence, vol. 3, no. 1, pp. 29-38, 2014.
[139] H. Xiong, J. J. Wu, J. Chen, "K-means clustering versus validation measures: a data-distribution perspective. IEEE Trans. Syst. Man Cybern. B Cybern. Vol. 39, no. 2, pp. 318–331, 2009.
[140] L. Kaufman, and P. J. Rousseeuw, "Clustering by means of Medoids," in Statistical Data Analysis Based on the L1 - Norm and Related Methods, North-Holland, pp. 405–416, 1987.
[141] A. K. Jain, M. N. Murty and P.J. Flynn, "Data Clustering: A Review," ACM Comput. Surveys, vol. 31, pp. 264-323, Sep. 1999.
[142] J.C. Gower, "A General Coefficient of Similarity and Some of Its Properties," Biometrics, vol. 27, no. 4, pp. 857-871, 1971.
[143] Z. X. Huang," Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.
[144] R. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proceedings of the 20th International Conference on Very Large Databases, Santiago, 1994.

[145] L.J.P. van der Maaten, “Accelerating t-SNE using Tree-Based Algorithms,” Journal of Machine Learning Research, vol. 15, pp. 3221-3245, 2014.
[146] L.J.P. van der Maaten and G.E. Hinton, “Visualizing High-Dimensional Data Using t-SNE,” Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.
[147] Swati Sahai and Anil Pahwa, “A Probabilistic Approach for Animal-Caused Outages in Overhead Distribution Systems”, 9th International Conference on Probabilistic Methods Applied to Power Systems, Stockholm, Sweden, June 11-15, 2016
[148] Odilon Faivre, Yann Le Herve, Pauline Jehl, et al, “Forecast of Faults During Heat Waves in A Medium Voltage Grid and Crisis Management”, CIRED Workshop 2016, Helsinki.
[149] Guangning Wu, Xuesong Ni, Zhenjie Song and Bo Gao, “Prediction for substation equipment failure rate based on improved grey combination model”. High Voltage Engineering, vol. 43, no. 7, pp. 2249-2255, 2017.
[150] Raffi Avo Sevlian, Yue Zhao, Ram Rajagopal, et al, “Outage Detection Using Load and Line Flow Measurements in Power Distribution Systems”, IEEE Transactions on Power Systems, vol. 33, no. 2, pp. 2053-2069, 2018
[151] Y. Wang, C. Chen, J. Wang, et al, “Research on resilience of power systems under natural disasters – A review”, IEEE Transactions on Power System, vol. 31, no. 2, pp. 1604-1613, 2016.
[152] R. Nateghi, S. Guikema, and S. M. Quiring, “Power outage estimation for tropical cyclones: Improved accuracy with simpler models”, Risk Analysis, vol. 34, no. 6, pp. 1069-1078, 2014.
[153] M. H. Wang, C. P. Hung, “Novel Grey Model for the Prediction of Trend of Dissolved Gases in Oil-filled Power Apparatus”, Electric Power Research, vol. 67, pp. 53-58, 2003.
[154] Yenhung Lin, Jeanshyan Wang and Pingfeng Pai, “A Grey Prediction Model with Factor Analysis Technique”, Journal of the Chinese Institute of Industrial Engineers, vol. 21, no. 6, pp. 535-542, 2004
[155] Z. X. Wang, Q. Li and L. L. Pei, “A seasonal GM(1,1) model for forecasting the electricity consumption of the primary economic sectors”, Energy, vol. 154, pp. 522-534, 2018.
[156] Song Ding, Keith W. Hipel and Yaoguo Dang, “Forecasting China’s Electricity Consumption Using A New Grey Prediction Model”, Energy, vol. 149, pp. 314-328, 2018
[157] Tianshan Gao and Bo Gao, “Failure rate prediction of substation equipment combined with grey linear regression combination model”, IEEE International Conference on High Voltage Engineering and Application (ICHVE), 2016, Chengdu, China.
[158] Holland J., Wang Z., Lv Z., et al, “Genetic algorithm and the optimal allocation of trials”, SIAM Journal of Computing, vol. 2, pp. 88-105, 1973
[159] Y. Liu, S. Lei and Y. Hou, “Overhead transmission line outage rate estimation under wind storms”, IEEJ Transactions on Electrical and Electronic Engineering, vol. 14, pp. 57-66.

[160] M. Yue, T. Toto, M. Jensen, et al, “A Bayesian approach-based outage prediction in electric utility systems using radar measurement data”, IEEE Transactions on Smart Grid, vol. 9, no. 6, 6149-6159, 2018
[161] P. Kankanala, S. Das and A. Pahwa, “Adaboost: an ensemble learning approach for estimating weather-related outages in distribution systems”, IEEE Transactions on Power Systems, vol. 29, no. 1, pp. 359-367, 2014
[162] A. Jaech, B. Zhang, M. Ostendorf, et al, “Real-time prediction of the duration of distribution system outages”, IEEE Transactions on Power Systems, vol. 34, no. 1, pp. 773-781, 2019
[163] N. Chawla, K. Bowyer, L. Hall, et al, “SMOTE: synthetic minority over-sampling technique”, Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002
[164] A. Bosisio, A. Berizzi, D. Le, et al, “Improving DTR assessment by means of PCA applied to wind data”, Electric Power System Research, vol. 172, pp. 193-200, 2019
[165] N. Singh, P. Singh and D. Bhagat, “A rule extraction approach from support vector machines for diagnosing hypertension among diabetes”, Expert Systems with Applications, vol. 130, pp. 188-205, 2019
[166] The Edison Foundation, Electricity Company Smart Meter Deployments: Foundation for A Smart Grid, December 2017.
[167] Berg Insight, Smart Metering in Europe, 2018.
[3] USmartConsumer Project, European Smart Metering Landscape Report, 2016.
[168] P. M. Quevedo, J. Contreras, A, Mazza, et al, “Reliability assessment of microgrids with local and mobile generation, time-dependent profiles and intraday reconfiguration”, IEEE Trans. Indus. Appl., vol. 54, no. 1, pp. 61-72, 2018.
[169] Y. Zhang, W. Z. Wu, C. H. Huang, “Steady-state analysis for isolated three-phase induction generator system with asymmetric loads”, Dianli Zidonghua Shebei/Electric Power Automation Equipment, vol. 37, no. 2, pp. 171-175, 2017.
[170] Sook-Chin Yip, Wooi-Nee Tan, ChiaKwang Tan, et al, “An Anomaly Detection framework for Identifying Energy Theft and Defective Meters in Smart Grids”, Electrical Power and Energy Systems, vol. 101, pp. 189-203, 2018.
[171] Tang Yijia and Gao Hang, “Anomaly Detection of Power Consumption based Waveform Feature Recognition”, the 11th International Conference on Computer Science and Education, August 23-25, 2016, Japan.
[172] A. Kewo, P. Manembu, X. Liu, P.S. Nielsen. Statistical Analysis for Factors Influencing Electricity Consumption at Regional Level. Proc. of IEEE 7th International Conference on Power and Energy (PECon), pp. 132-137, 2018.
[173] Hanxiang Ding, Kun Ding, Jingwei Zhang, et al, “Local Outlier Factor-based Fault Detection and Evaluation of Photovoltaic System”, Solar Energy, vol. 164, pp. 139-148, 2018.
[174] Peter de Wilde, Wei Tan, “Predicting the performance of an office under climate change: a study of metrics, sensitivity and zonal resolution”, Energy and Buildings 2010;42:1674–84.

[175] Gemma Manache and Charles Melching, “Identification of Reliable Regression and Correlation-based Sensitivity Measures for Importance Ranking of Water-Quality Model Parameters”, Environmental Modeling and Software, vol. 23, pp. 549-562, 2008.
[176] Sayanti Mukherjee and Roshanak Nateghi, “Climate Sensitivity of End-use Electricity Consumption in the Built Environment: An Application to the State of Florida, United States”, Energy, vol. 128, pp. 688-700, 2017.
[177] Wei Tian, “A Review of Sensitivity Analysis Methods in Building Energy Analysis”, Renewable and Sustainable Energy Reviews, vol. 20, pp. 411-419, 2013.
[178] Wei Tian, Pieter de Wide, “Uncertainty and Sensitivity Analysis of Building Performance using Probabilistic Climate Projections: A UK Case Study”, Automation in Construction, vol. 20, pp. 1096-1109, 2011.
[179] Leo Breiman, “Random forests”, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[180] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, et al, “LOF: Identifying Density-based Local Outliers”, the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, USA — May 15 - 18, 2000.
[181] Organization for Security and Co-operation in Europe (OSCE), Protecting Electricity Networks from Natural Hazards, 2016, http://www.osce.org/secretariat/242651?download=true, (available on line on 30th March 2017).
[182] M. P. Panteli, D. N. Trakas, P. Mancarella, et al, “Boosting the power grid resilience to extreme weather events using defensive islanding”, IEEE Transactions on Smart Grid, vol. 7, no. 6, pp. 2913-2922, Nov. 2016.
[183] A. C. Reilly, G. L. Tonn, C. Zhai, et al, “Hurricanes and power system reliability – The effects of individual decisions and system-level hardening”, Proceedings of the IEEE, vol. 105, no. 7, pp. 1429-1442, Jul. 2017.
[184] EURELECTRIC, “Power Distribution in Europe Facts and Figures”. [online]. Available: https://www.eurelectric.org/news/eurelectric-publishes-facts-and-figures-on-power-distribution-in-europe
[185] Department of Economic and Social Affairs, United Nations, “2018 Revision of World Urbanization Prospects”. [online]. Available: https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html
[186] M. Panteli, C. Pickering and S. Wilkinson, et al., “Power System Resilience to Extreme Weather: Fragility Model, Probability Impact Assessment and Adaptation Measures”, IEEE Trans. Power System, vol. 32, pp. 3747-3757, 2017.
[187] F. Cadini, G. L. Agliardi, E. Zio, “A modeling and simulation framework for the reliability/availability assessment of a power transmission grid subject to cascading failures under extreme weather conditions”, Applied Energy, vol. 185, pp. 267-279, 2017.

[188] M. Panteli and P. Mancarella, “Influence of extreme weather and climate change on the resilience of power system: Impacts and possible mitigation strategies”, Electric Power System Research, vol. 127, pp. 259-270, 2015.
[189] Y. Zhang, A. Mazza, E. Bompard, et al, “Data-driven feature description of heat wave effect on distribution system”, IEEE Conference on Powertech, Milan, Italy, 2019.
[190] M. Panteli, D. N. Trakas, P. Mancarella, et al, “Power system resilience assessment: Hardening and smart operational enhancement strategies”, Proceedings of the IEEE, vol. 105, no. 7, pp. 1202-1213.
[191] M. Panteli, C. Pokering, S. Wilkinson, et al, “Power system resilience to extreme weather: Fragility modeling, probabilistic impact assessment and adaptation measures”, IEEE Transactions on power systems, vol. 32, no. 5, pp. 3747-3757, 2017.
[192] S. Espinoza, M. Panteli, P. Mancarella, et al, “Multi-phase assessment and adaptation of power systems resilience to natural hazards”, Electric Power Systems Research, vol. 136, pp. 352-361, 2016.
[193] ENTSO-E, “Guideline for Cost Benefit Analysis of Grid Development Projects”, 2015, https://www.entsoe.eu/fileadmin/user_upload/_library/events/Workshops/CBA/121119_CBA_introduction.pdf, (available on line on 30th March 2017).
[194] World Energy Council, “The road to resilience – managing and financing extreme weather events”, 2015.
[195] M. Smid, S. Russo, A. C. Costa, et al, “Ranking European capitals by exposure to heat waves and cold waves”, Urban Climate, vol. 27, pp. 388-402, 2019