PHOTOVOLTAIC SYSTEMS CLASSIFICATION AND SIZING BASED ON THE HISTORICAL POWER FLOW DATA

by

XIAOTONG WANG

B.A.Sc., The University of British Columbia, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

Photovoltaic Systems Classification and Sizing Based on the Historical Power Flow Data

submitted by Xiaotong Wang in partial fulfillment of the requirements for

the degree of Master of Applied Science

in Electrical and Computer Engineering

Examining Committee:

William Dunford
Supervisor

Rabab Ward
Supervisory Committee Member

Y. Christine Chen
Supervisory Committee Member

Abstract

As the development of sensor and data storage technology, more data have become available for analysis. A commercial database stores the power flow data for more than 4000 Photovoltaic systems. Various optimization methods have been researched for reducing the PV system's cost by sizing each system component appropriately. For most of the existing optimization methods, they focus on the computational efficiency, system modelling or data availability. The disadvantage is they always assume the system's type and operation strategy are known. However, in the given database, the system's type is unlabeled.

This thesis proposes a method for sizing PV systems based on their historical power flow data stored in the multivariate time-series format. The method is presented in three consecutive steps. In the first step, a validation rule is applied to filter out the problematic PV systems. The systems whose battery monitor is incorrectly installed can also be detected by the Gaussian Mixture Model (GMM) method, and the related data can be fixed afterward. In the second step, seven features are determined to differentiate the PV systems. The GMM method is applied to cluster the PV system based on the proposed features, so we can identify the system's type through visualizing the classification results. Once we know the system type, in the last step, the PV system is modelled mathematically. The Artificial Bee Colony (ABC) method is implemented based on the system model, historical data, and operation strategy to determine the optimal size of each system component. As an example, a stand-alone system is chosen to demonstrate the process that determines the sizes of the PV panel, diesel generator, and battery bank.

Lay Summary

PV energy is an environmentally friendly source and the application of PV system may remit the emission of greenhouse gas and air pollution. Moreover, it becomes an ideal energy source for the communities in the remote areas, where the grid extension is not economical. However, its cost still concerns potential users, and especially, its initial capital cost is usually higher than other types of system. To make the PV energy more economical, this thesis proposes a procedure to minimize the overall cost of PV systems based on a database that records all systems' historical power flow data. It first extracts a few features from the time-series data to obtain each system's operation behaviour. Then, based on the selected features, similar systems are grouped. Therefore, cost functions can be formed for each group, and they are converted to an optimization problem to determine the most economical size of PV array, battery bank capacity and the power rating of a back-up diesel generator.

Preface

This study is based on a commercial database that monitors PV systems' operation and records the power flow data. Mr. Masautso Sau Ngosi, Mr. Leon Hailstones, and Mr. Trevor Monk explained the database structure and the possible topologies of PV systems. The thesis proposes a procedure that reduces the system's cost by finding the sizes of system components. I am the major contributor and responsible for reviewing the relevant studies, data cleaning and correction, methods selection, designing features, programming, and analyzing classification results. The project was also performed under the MITACS Accelerate Grant "Microgrid's Performance Modeling & Optimization Method Based on Data Mining & Artificial Intelligence." My research supervisor, Dr. William Dunford, has provided feedback and comments throughout the process of performing the research and writing the thesis. Dr. Rabab Ward also helped me in my thesis writing.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABC Artificial Bee Colony

AC Alternating Current

AGS Automatic Generator Starter

ANN Artificial Neural Network

BIC Bayesian Information Criterion

DC Direct Current

EA Evolutionary Algorithm

GMM Gaussian Mixture Model

HMM Hidden Markov Models

ICL Integrated Completed Likelihood

LLP Loss of Load Probability

MCN Maximum Cycling Number

MPPT Maximum Power Point

PCA Principal Component Analysis

PSO Particle Swarm Optimization

PV Photovoltaic

SCP System Control Panel

SoC State of Charge

UPS Uninterruptible Power Supply

Acknowledgements

First, I would like to thank my academic supervisor, Dr. William Dunford for his advice and guidance throughout this research. I would also like to thank Mr. Masautso Sau Ngosi who founded the subject through the MITACS Accelerate Grant.

I am particularly grateful to Mr. Leon Hailstones and Mr. Trevor Monk who took the time to explain the commercial database and devices in a PV system. Special thanks to Ran Liao, Yi-Ran Zhang, Jia-Xi Gao, Zhi Qu, Xue-Jun Ji, Zheng Hong, Zhen-Dong Cao, Xin-Yuan Zhang and Chang Ge, for proofreading my writing and providing invaluable feedback.

I am very happy to have the chance to study here at UBC in Canada, the experience and interactions with the professors, staff, and students have been wonderful. I thank my parents, my girlfriend and my family friend, Kui Wang for their unwavering support throughout my years of education. Their care, love, encouragement, financial support and company have carried me through the toughest of time.

Chapter 1: Introduction

Solar energy is clean and free from carbon dioxide emission during operation. It is an ideal alternative energy source that has been widely applied in various applications in recent years, especially for the communities in the rural area. However, the cost of a photovoltaic (PV) system remains the primary concern for potential customers. In an effort to reduce consumers' year- round cost, researchers have been investigating ways to determine the sizes of different components in the PV system. This thesis will focus on the capacities of batteries, and the power ratings of diesel generators and PV panels. A private company has provided us a commercial database that contains information of over 4000 PV systems implemented around the world, including the time-series power flow data being analyzed in this study. From analyzing the given database, this thesis proposes a process to improve the sizing of the existing devices in the PV systems and suggests modification in the database for future analysis.

1.1 Photovoltaic Generation

The global demand for electricity has been increasing drastically for the past decades. The consumption of fossil fuel causes a significant negative impact on the environment such as air pollution, global warming, climate change, etc. Therefore, renewable energy has been attracting a large amount of attention in the recent two decades [1]. Among all the alternative ways to harvest energy, PV has been more widely adopted. Studies have shown that there are many opportunities to implement profitable PV systems which would also reduce CO_2 emission [2] and benefit rural and remote communities[3], [4]. Compared to a PV system, diesel

generator-based systems have higher maintenance and operation costs [5]. In some remote areas, diesel is very expensive, and new extensions of the power grid is highly uneconomical [6]. In these cases, a stand-alone PV system or a micro-grid driven by PV generation will be more economical [7]. Although solar energy has many benefits over fossil fuel, the capital cost of a PV system is still high. This remains the leading limiting factor of the PV system implementation in rural areas [8]. In addition, PV has an uncertain and intermittent nature like other types of renewable energy. An energy storage system, such as a battery bank, can improve the stability and reliability for the consumer, but would require even more capital investment, increasing economic feasibility concerns [9].

1.2 The Given Commercial Database and The Commercial PV Systems

This research is based on data stored in a given commercial database collected through PV-system solutions provided by owners. The PV-system product line includes the inverter, maximum power point tracker (MPPT), automatic generator starter (AGS), system control panel (SCP), communication box (ComBox), battery monitor, and user interface through the internet. Customers who acquire the PV-system product line would need to procure battery banks, a diesel generator and PV panels from a third-party manufacturer. shows the layout of a fully extended system. In the system, devices are connected through an AC bus, a DC bus, and a communication bus. The arrows in the diagram show the possible directions of power flow. The functions of each component are described below.

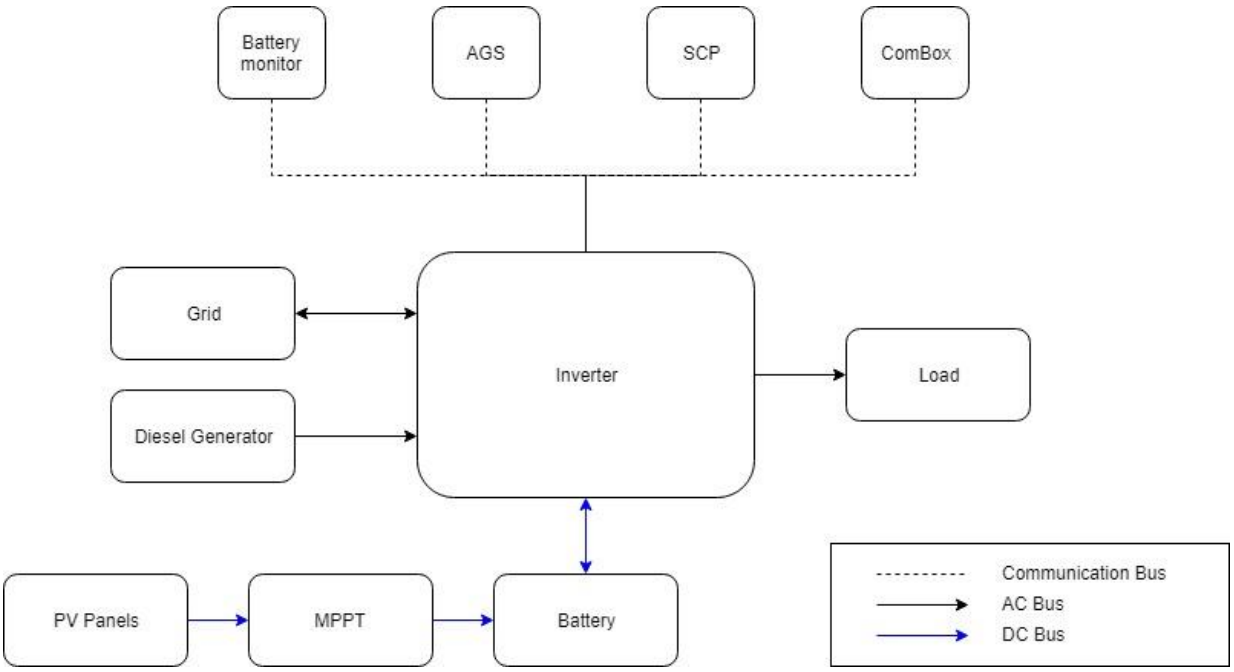

Figure 1.1 Block diagram of a fully extended PV system in the given database

1) Inverter: The inverter can convert DC to AC or AC to DC. It can also control the amount of energy drawn from each source and measure the amount of energy transmitted. The inverter acts as an interface between the DC bus, grid, backup source, and load.
2) Grid: The grid can provide and intake AC power, usually considered reliable and has no limitation on power rating. Not all systems have a grid connection and some grids are not reliable.
3) Diesel generator: The diesel generator consumes diesel and generates stable AC energy. The power rating limits its output. In a PV system, it can act as either the primary energy source or the backup energy source. The diesel generator is usually installed when the grid is not accessible.

4) PV panels: The PV panels generate DC energy under solar radiation through the photovoltaic effect. Their output power relates to the panel's power rating, the strength of the radiation and the panel's orientation.
5) MPPT: It is essentially a DC/DC converter which controls (usually by maximizing) the energy from the PV panels according to the panel's V-I characteristic curve.
6) Battery: An electrical energy storage device. Most studied sites adopt lead-acid batteries and a few of sites install lithium-ion batteries.
7) Load: A load in the system can be anything that consumes energy. In most cases, the primary purpose of a system is to satisfy the load's demand. However, some systems do not have a load, or their loads are minimal, in which case most of the generate energy will be injected to the grid.
8) Battery monitor: It measures current, voltage, temperature, state of charge (SoC), average cycling depth, number of cycles, and remaining time of a lead acid battery bank. It can connect with a ComBox through a communication bus to update battery information in the database.
9) AGS: The AGS can start or stop a generator based on the shared status information over the communication bus. It responds to a set of programmable requirements such as battery voltage, battery SoC and grid power.
10) SCP: The SCP is a user interface that informs users of the system's status, warnings, and errors. The displayed information includes the battery's SoC, temperature, MPPT's harvest and the currents on each bus, etc. A user can also set configurations and parameters through the SCP.

11) ComBox: It is a communication device integrated with a web server. It uploads the system's status to the web server so users can use mobile devices or PCs to access the system's status. Comparing to an SCP, the ComBox enables remote monitoring and configuration. It also provides a more complexed visualization tool and analysis.

The integrated web server is one of the system's features. It can collect data from each system. The information we will analyze comes from this collection. There are three types of data based on how they are updated:

1) Manually updated data: It includes the devices' configuration, site name, address, battery size, and panel size, etc. They are entered manually by the user.
2) Dynamic data: Dynamic data is measured and updated to the database automatically, but the database only stores the latest data. It includes the battery's temperature, SoC, voltage, etc.
3) Historical data: In the database, only the system's power flow have historical data. In each site, seven types of power flow are available, including PV generation, load consumption, grid input, grid output, battery charged energy, battery discharged energy and generator's generation. All historical data are stored once every ten minutes. Figure 1.2 gives an example of a PV system’s daily power flow profile. On a given day, the system only records three types of power flow; PV input, grid output and load output. The power flow profile helps the user to understand the system’s behaviour. Figure 1.2 shows that the system generates solar energy and injects most of the energy to the connected grid.

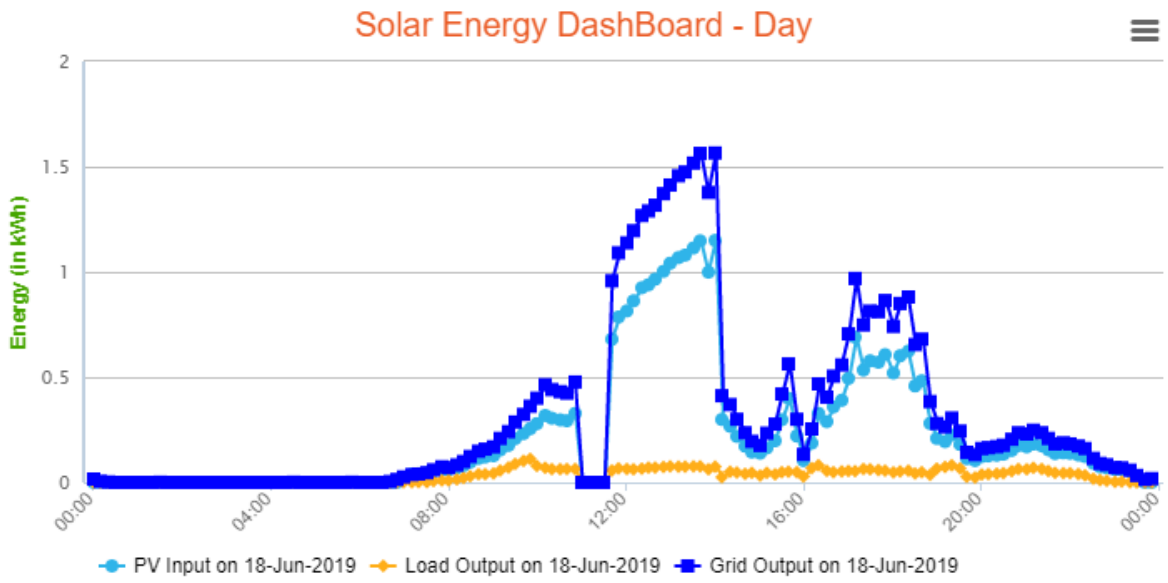

Figure 1.2 Example of a system's daily power flow profile

1.3 Application of Commercial and Residential PV Systems

The fully extended system shown in includes all possible devices, but most PV systems only contain some devices. The PV systems in the given database are for various applications, a few possibilities are described below.

1.3.1 Off-Grid Solar Power Systems

For off-grid residences, PV is usually the primary power source for their load. The battery bank stores excessive power generated from PV panels during sunny periods for later usage. It's usually combined with a diesel generator as a backup source in case the battery runs out of energy. Or the diesel generator can be the primary energy source. In this case, the adoption

of PV as secondary generation shortens the usage time of diesel generator and reduces the cost of fuel.

1.3.2 Net Metering Systems

The net metering systems aim to maximize the yield in power generation and the return of investment. In this type of application, the MPPT is directly connected to the inverter, and the battery is not included. If there is a load, power will be supplied by the PV system and/ or the power grid. Excessive power (or where there's no load, all power) generated through the PV system are sold back to the grid.

1.3.3 Storage and Backup Systems

The grid is considered the primary source in these systems and the power generated through the PV system is stored in battery banks and only used for backup purposes. The inverter converts DC power from the battery to AC and seamlessly connects it to the load when the grid is not available.

1.3.4 Self-Consumption with Storage Systems

A self-consumption with storage system usually connects to the grid and has a DC-coupled PV generation with storage. The application prioritizes energy usage from the DC bus until the battery SoC drops to a preset threshold, then switches to purchasing power from the

grid. This allows the battery bank to also act as a backup source when power from the grid is not available. When there is sufficient radiation, excessive power generated from PV panels charges the battery.

1.4 Research Objectives

To promote the use of solar power systems by satisfying demands and minimizing system cost, it is wise to design for appropriate sizes for each component in the system. The commercial database has been collecting power flow data from their PV systems for years. Even though there's a massive amount of data collected, the data from different types of PV systems are mixed together and need to be classified and validated prior to further analysis. In this study, data analysis, visualization and simulation are implemented in MATLAB and R. The overall goal is divided into the following three objectives:

1.4.1 Objective 1: Data Cleaning for the Database

After looking through the data provided, some of the values appeared questionable. Poor data quality can be due to many reasons. The Gaussian mixture models (GMM) method is implemented in this thesis to identify erroneous connections of the battery monitor to the system, and to correct the data accordingly. The GMM classification method tolerates outliers, has fewer parameters to be configured and allows clusters to overlap. For errors with unknown

causes, validation rules are applied to identify the valid data from the database. The invalid data are excluded from classification and analysis in Chapter 3.

1.4.2 Objective 2: Identifying System's Application Based on Power Flow Data

As discussed previously, PV systems can be applied for different applications. Although all PV system users update their site status, they do not specify the application of their site. The system's application is important information because a system cannot be optimized without prior knowledge of the system's purpose of operation. From their power flow, it is possible to identify the primary source, operation strategy and the application. Both poor data quality and unique applications in systems can cause outliers in the database. The classification method should identify outliers.

1.4.3 Objective 3: Sizing Panels, Battery Bank and Diesel Generator to Minimize System's Cost

Once the application is determined, it is possible to form a function representing its year-round cost. Various types of costs may be associated with each component in a system. All costs should be formularized and summarized. The optimization purpose is to minimize the year-round cost with respect to constraints. This study proposes two methods: the first method is based on the given power flow data and optimizes the size of the battery. The second method is versatile, and sizes several components simultaneously. However, it needs additional information that the database does not provide.

Chapter 2: Data Cleaning

The given database tracks more than 4000 PV systems, but only a portion of the systems directly monitor the battery's charging and discharging power flow. Our analysis only considers these PV systems because their data are more reliable. The raw data has many other quality issues, which will be discussed in this chapter. In order to filter out unreliable data and carry further analysis, a validation rule is proposed. In addition, a method is proposed to detect the incorrect installation of the battery monitor and correct the data.

2.1 Problems of Data Quality

Each site has up to seven sets of historical data for analysis. However, there are various problems in the data, which will affect our analysis. Therefore, before the data are thoroughly analyzed, it is necessary to clean up the data itself. Different types of quality issues are treated differently. The data will be considered for analysis if their errors can be corrected, otherwise they will not be included in the analysis. The main issues with the data are discussed in sections 2.1.1 to 2.1.5.

2.1.1 Dummy and Deleted Sites

This database contains many dummy sites and deleted sites. These sites do not record valid energy flow information, so they are not useful for research. Deleted sites often have various problems, such as missing records, short history, or duplicate registrations. There is no data recorded at all in the example shown in figure 2.1. Therefore, these systems are excluded from further analysis.

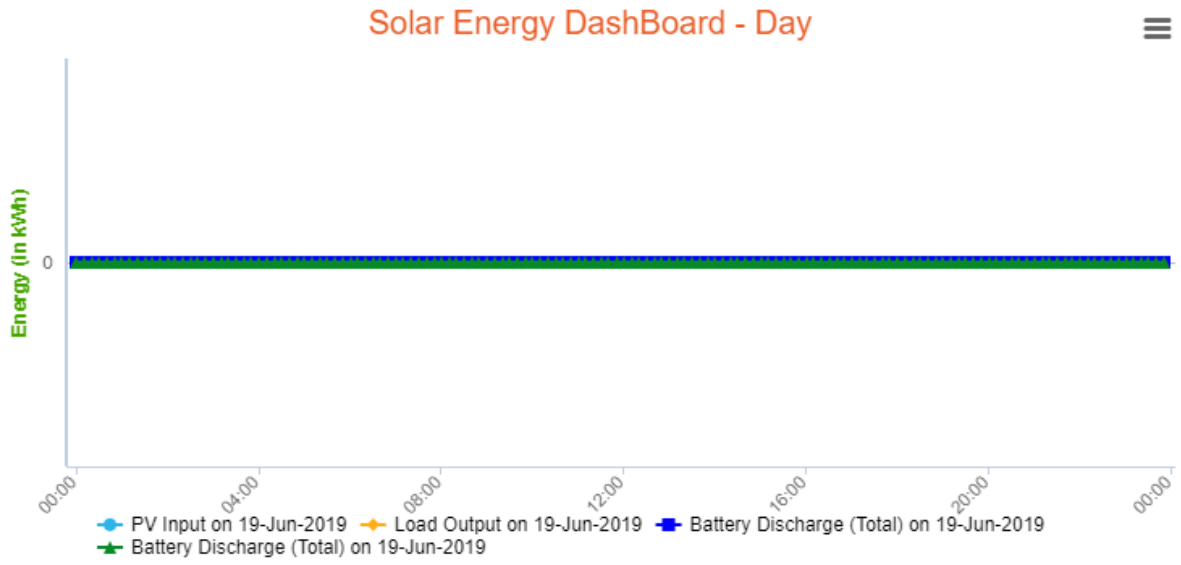

Figure 2.1 An example of dummy system's daily power flow profile

2.1.2 Missing Data

The database provided is missing data. One situation is that a particular type of data is absent. Sometimes a type of energy flow exists within the system but is not recorded. The typical

case is the lack of PV power generation records. In this case, we cannot understand the behaviours of the site at all, and we cannot be sure what energy sources exist at the site. According to Figure 2.2, the battery was charged, and the load consumed energy at noon, but no energy was supplied through any means. The PV monitor is likely broken, but other potential causes (eg. an unmonitored source in the system, a broken generator monitor) cannot be excluded.

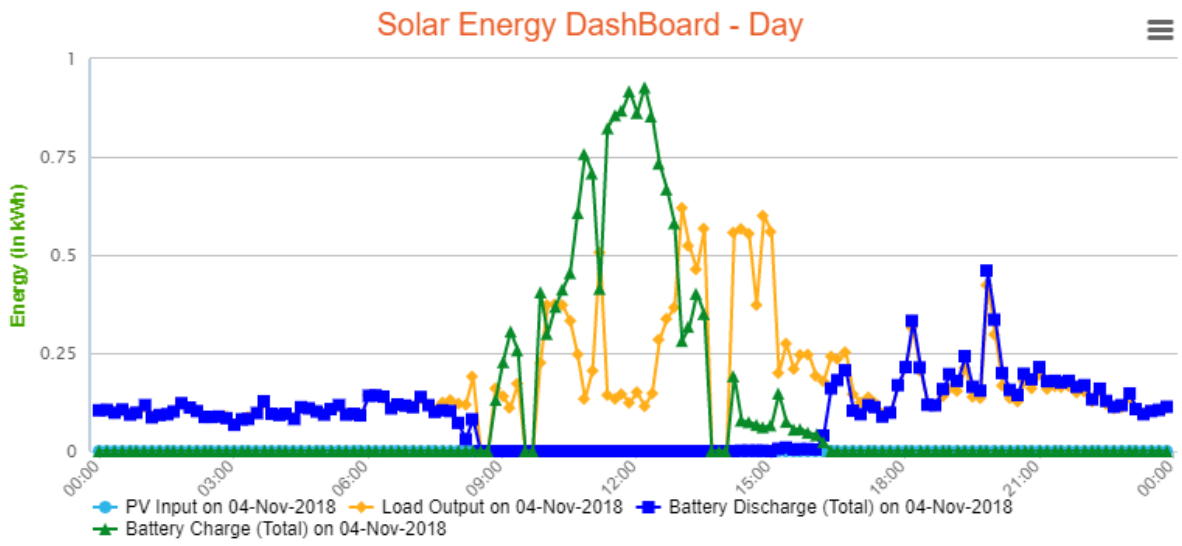

Figure 2.2 An example of missing data

Another situation is where the data is missing during a specific period. For example, no data was recorded for a few months, or the system only updates data once a week. The data collected in these situations can be very misleading. However, some sites only lose a few time steps of data, this type of loss will not affect the analysis of their overall behaviour. Figure 2.2 is an example of this case. It misses a few data points at noon, but the user can still understand the system's behaviour.

2.1.3 Outliers

Outliers may imply special events or may be caused by genuine system errors. In a grid-tie system, a battery bank may operate as an alternative source. It usually does not supply energy to loads. However, when the grid fails the battery starts to discharge, and its data become outliers. In another example, some data have no apparent explicable cause and are well outside the normal range, as shown in Figure 2.3.

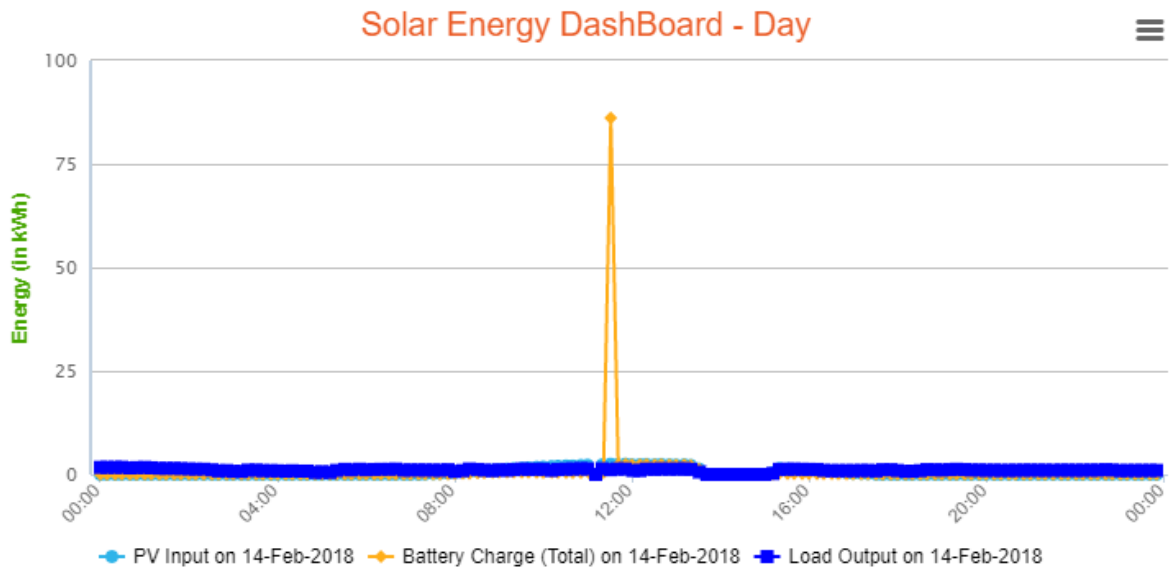

Figure 2.3 An extreme value in the record

2.1.4 Scaled Measurement

Incorrect upscaling or downscaling can occur but is not easy to detect. It is usually due to a sensor installation error or wrong configurations, but the measured value is proportional to the

real value. In Figure 2.4, the generator input data is proportional to but greater than the battery charge data at night. There is no load or grid output during these hours. So theoretically the generator input should be equal to the battery charge data. We can deduce that at least one set of the measurements is scaled, but it is not possible to determine if either are valid.

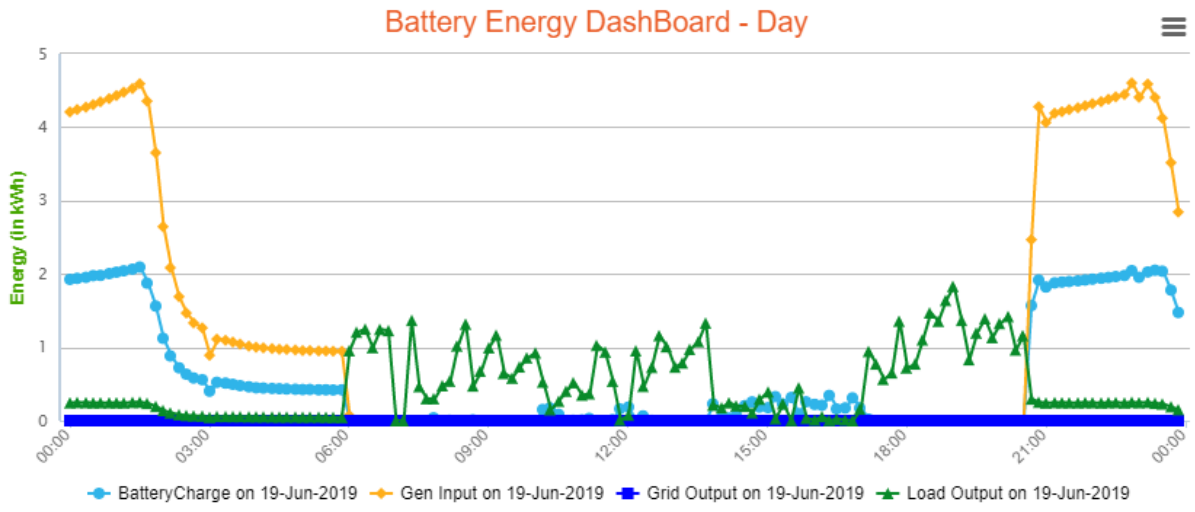

Figure 2.4 Scaled measurement in the database

2.1.5 Device Installation Error

The system has a default topology for measuring up to seven types of power flow. If a local installer installed the sensors or devices incorrectly, the data would not be describing what was intended to be measured. In Figure 2.5 the grid input and the load output change synchronously and drop to zero at noon, but the grid output increased significantly. A reasonable guess is that the PV panel coupled on the AC bus and the load output is measured incorrectly.

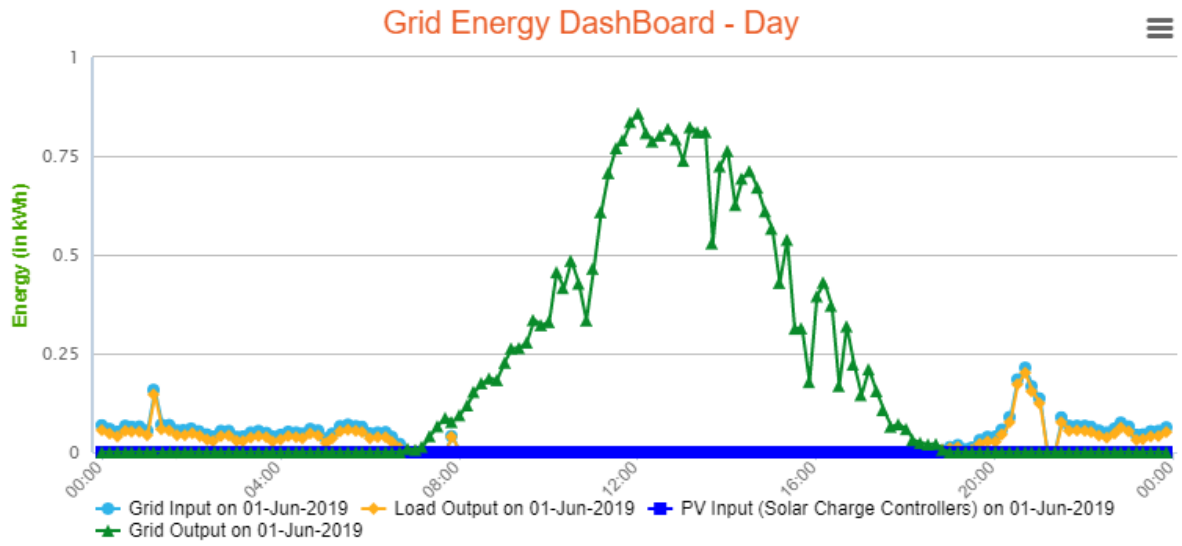

Figure 2.5 Wrong measurement due to incorrect device installation

A prevalent mistake is that the battery monitors were installed at the wrong location or in the wrong direction. This type of error can be detected and corrected. The correction procedure is described in Section 2.3.

2.2 Data Validation

The database contains many quality problems which cause noise during data processing and analysis. Useful information may be undetectable under the noise, so the invalid data should be either removed or corrected. One straightforward method to validate data is based on the law of conservation of energy.

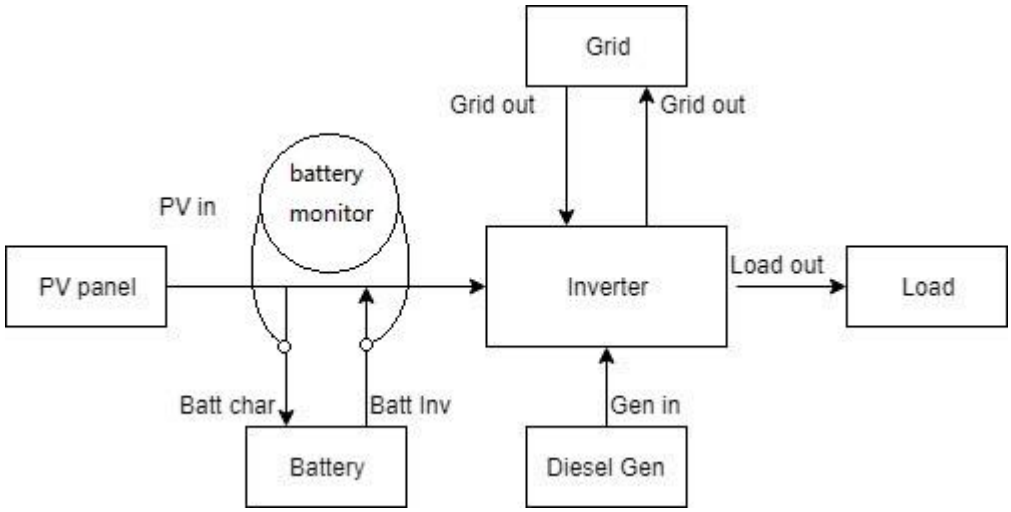

Figure 2.6 Seven measured power flow in the commercial and residential PV system.

As the typology shown in figure 2.6, the inverter is a node in the system. According to the law of conservation of energy or Kirchhoff's current rule. The energy going into the node is equal to the energy out of the node:

$$E_{In}[t] = E_{\text{out}}[t] + E_{loss}[t], \tag{2.1}$$

where

$$E_{In}[t] = E_{PVIn}[t] + E_{BattInv}[t] + E_{GenIn}[t] + E_{GridIn}[t] \tag{2.2}$$

$$E_{\text{out}}[t] = E_{load}[t] + E_{GridOut}[t] + E_{BattChar}[t] \tag{2.3}$$

$E_{PVIn}, E_{BattInv}, E_{GenIn}, E_{GridIn}, E_{load}, E_{GridOut}, E_{BattChar}$ are the seven power flows which have been recorded in the database. E_{loss} includes DC/AC conversion losses, and the energy for controller and communications. When the converted energy is large, the E_{loss} approximately equal to $(1-\eta)(E_{load} + E_{GridOut})$. The conversion efficiency is about 97%. Therefore, a site should satisfy the following approximation.

$$\frac{E_{out}[t]}{E_{in}[t]} \approx \eta \tag{2.4}$$

Based on the data in the recent year, the injected energy and extracted energy in each site can be calculated. The calculated result can be plotted on the following scatter plot.

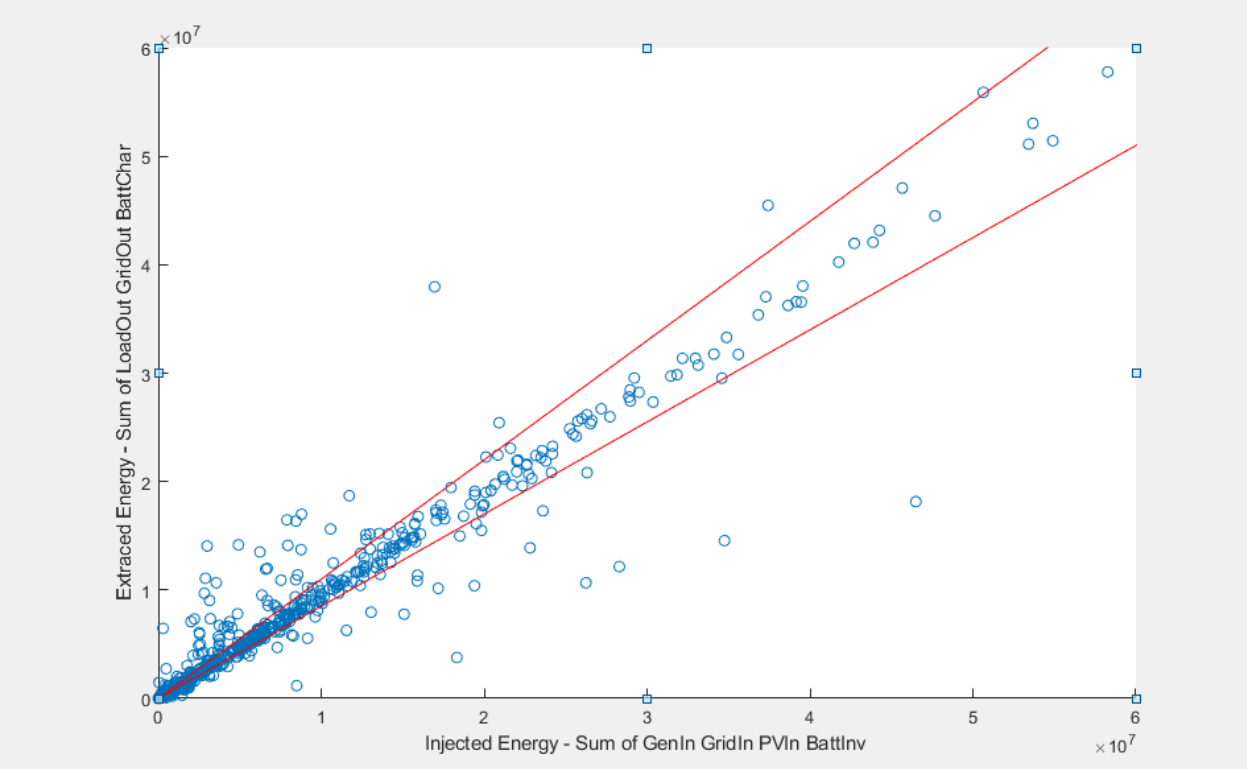

Figure 2.7 Data validation based on the law of conservation of energy: comparison of extracted energy and injected energy.

By observing the plot above, we can find that most sites are within the acceptable range which is represented by two guidelines, especially as the injected/ extracted energy increases. The slope of the line should be the conversion efficiency of the inverter. Because the approximation is based on some assumptions and the measurements may not be accurate, I leave a margin for all sites, any site satisfying the following inequation will be trusted, and their data are valid data.

$$0.85 * E_{In} < E_{\text{out}} < 1.1 * E_{In} \tag{2.5}$$

There are many reasons causing energy imbalance as discussed in Section 2.1. Any data quality issue may break the conservation. Since the database does not provide redundant measurements, it is impossible to correct those quality issues. Therefore, all sites which do not satisfy the inequality will not be considered in further analysis.

2.3 Battery Charge and Discharge Data Correction

As mentioned in Section 2.1.5, the battery monitor may be installed incorrectly. The default topology of the PV system is shown in Figure 2.6. The battery monitor should measure the power flow going into and out of the battery. However, the battery monitor was connected to measure the power flow out of the MPPT and the power flow into the inverter. These data were labeled as charged energy and discharged energy respectively on many sites. The typology of such sites is shown in Figure 2.8.

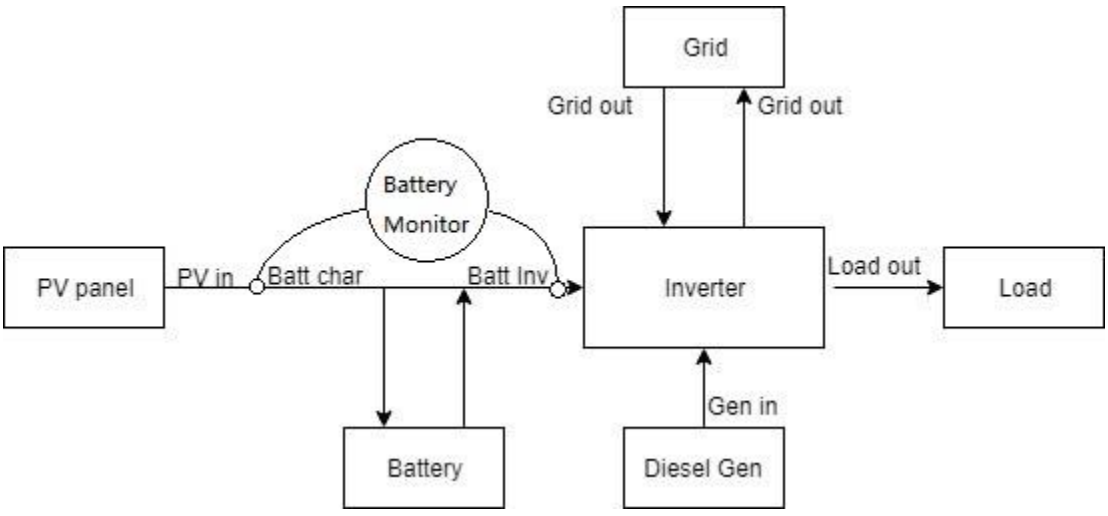

Figure 2.8 Battery monitor wrongly installed in the commercial and residential PV system and measuring incorrect battery charge and discharge power flow.

Even though the law of energy conservation is applied to validate the dataset, the error mentioned above will not be detected. If we treat the inverter as a node again we get the following equation for the typology shown in Figure 2.8:

$$E_{BattInv}[t] + E_{GenIn}[t] = E_{load}[t] + E_{GridOut}[t] + E_{loss}[t] \quad (2.6)$$

Due to the wrong sensor installation location,

$$E_{PVIn}[t] = E_{BattChar}[t] \quad (2.7)$$

Adding both sides of the equations 2.6 and 2.7, according to the definition in equations 2.2 and 2.3 we get,

$$E_{In}[t] = E_{\text{out}}[t] + E_{loss}[t] \quad (2.8)$$

Equation 2.8 is exactly the same as the validation rule (equation 2.1.) This means that the validation process cannot detect data collected through incorrectly installed battery monitors.

To detect the error and fix it, the GMM classification is implemented by thinking the difference between the normal systems and the systems whose battery monitor is incorrectly installed. Therefore, the following three features are selected, and they are explained in detail.

1) The ratio of battery charged energy to PV generation

When the battery charge sensor is incorrectly installed, it measures the PV generation rather than charged energy. Therefore, the ratio of the two will be close to 1. In an efficient PV system, solar energy is preferably consumed directly rather than stored in a battery, because the

charging process causes energy losses and battery degradation. When the sensors are installed correctly, the ratio of battery charged energy to PV generation is low.

2) The ratio of micro-cycling energy to battery charged energy

A battery cannot charge and discharge at the same moment. Usually, it charges when the system generates spare energy and discharges when the load demands. During a short period, a battery may switch between the charge mode and discharge mode frequently. The behaviour is called micro-cycling. This behaviour usually happens during sunset and sunrise. During these periods, the demand for power is roughly equal to supplied power. Any perturbation will be buffered by the battery through micro-cycling. The amount of energy cycled during the micro-cycling is small. It only takes a small portion of the charged energy. Therefore, when all sensors are installed correctly, the ratio should be low. However, because we prefer the PV energy to be consumed directly, the ratio of PV generation to the inverted energy will be close to 1.

$$E_{cycled}[t] = \min(E_{BattChar}[t], E_{BattInv}[t]) \tag{2.9}$$

3) The ratio of battery charged energy to battery discharged energy during the day

During the daytime, PV usually generates energy to support loads, and the excess energy is stored into the battery. Since the PV is the primary source, the battery is unlikely to discharge. The ratio of discharged energy to charged energy should be close to zero.

As discussed previously, PV generation is equal to the measured battery charge energy in systems with incorrectly installed sensors. In the daytime, PV tends to supply energy to the loads, and the excessive energy is stored in the battery. Therefore, the inverted energy is less than the PV generation. The site wrongly measures PV generation as the charged energy, the amount of cycled energy is equal to the charged energy or discharged energy, whichever has a lesser value; the inverted energy is less than PV generation over time. Hence the cycled energy equals the measured discharged energy when the sensor is installed incorrectly, which implies the following equation. In other words, as noted in Table 2.1, Feature#3 equals Feature#2.

$$\frac{Cycled\ Energy}{PV\ Generation} = \frac{Discharged\ Energy}{Charged\ Energy} \tag{2.10}$$

	Normal systems	Systems with incorrectly installed battery monitor
Feature #1	Much lower than 1	Close to 1
Feature #2	Close to 0	Greater than 0
Feature #3	Close to 0	Equals feature #2

Table 2.1 Expected features' value with respect to normal systems and systems with incorrectly installed battery monitor.

The three features for all sites are calculated based on the most recent year-round data. We set the amount of cluster to be six, and the GMM method generates the classification result as shown in Figure 2.9. The six categories are shown with different colours and symbols.

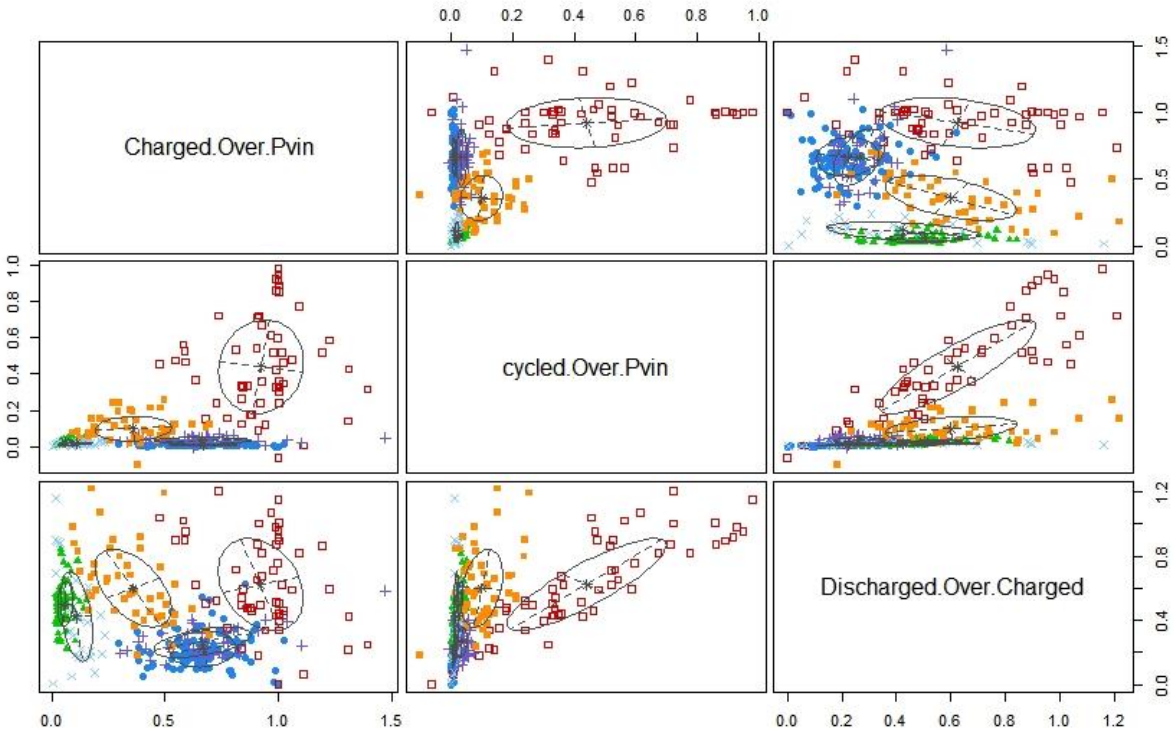

Figure 2.9 GMM classification result for identifying incorrectly installed battery monitor.

The cluster consisting of the red hollow squares should be emphasized. Sites in this category are suspected of installing their battery monitors incorrectly. Their features' value meet the expectations listed in Table 2.1. On the contrary, other clusters tend to have small values in the three features. They are classified into a different category according to their operation mode or behaviour. Details will be discussed in later chapters.

Once the faulty systems are detected, their data can be corrected in the following steps. When the PV generation is greater than the demand, the excessive energy goes into the battery.

$$E_{ActureBattChar} = \max(0,\ E_{MeasuredBattChar} - E_{MeasuredBattInv}) \tag{2.11}$$

When the PV generation is insufficient, the battery discharges to meet the demand.

$$E_{AcutureBattInv} = \max(0,\ E_{MeasuredBattInv} - E_{MeasuredBattChar}) \quad (2.12)$$

The errors in the database due to wrong installations is corrected. But it should be noted that their charged energy and discharged energy will not be positive at the same time step, which means we lose the information about their micro-cycling.

2.4 Conclusion

The database has many quality issues. To use accurate data, we only consider sites with a battery monitor. In 1168 sites with the battery monitor, 487 sites are either dummy sites or deleted. After applying the law of conservation of energy, 431 sites are valid. Within these sites, 33 sites do not have PV generation and some sites record less than 50 days' data. Finally, we keep only consider 358 sites for further analysis.

There are many possibilities for the invalid data and we only have limited information about each system, so it is very difficult to know the specific reasons that render the data invalid. Without knowing the cause, we cannot correct the wrong data. I excluded the invalid data because they behave like noise during the analysis and clustering process. They do not provided any helpful information and masks useful details.

Three features are selected to identify sites with battery monitors installed incorrectly. 55 faulty systems are found in the 358 sites. The verification process is done manually. Five systems on the cluster boundary and four systems at the center of the cluster are verified. The

four systems at the center of the cluster are correctly identified, but three out of five systems on the classification boundary are classified incorrectly.

Chapter 3: Site Classification

As discussed in Section 0, the PV system can operate differently to achieve various purposes. In the previous chapter, the data is validated and corrected. However, different types of systems are mixed in the database. It is impossible to optimize different systems in the same method because different PV systems are operated differently. Therefore, it is necessary to classify the PV systems according to their behaviours. Then, their operation purposes can be determined, and their component size can be optimized correspondingly. In this chapter, classification methods for multivariate time-series data are reviewed and feature selection are explained. Finally, the proposed method is implemented to classify the PV systems type.

3.1 Classification Methods for Multivariate Time-Series Data

Unsupervised classification organizes data that do not have class information into homogenous groups where the within-group-object similarity is maximized, and the between-group-object similarity is minimized [10], [11], [12]. Classic classification methods, such as hierarchical method, density-based method, grid-based method, etc. are based on data described with static features [10], [12]. However, in recent decades, more data is stored in the time-series format due to the development of sensing, storage and processor technologies [11], [13]. In the real-word, many applications are dynamical, and their data are stored in the time-series format such as sale data, waveform, robotic status, ecology data and so on [10], [11], [14], [15], [16], [17].

Compared to the static features, the time-series data are in much higher dimensions. Besides the complexity of computation, it is also challenging to compare the similarity of two time-series data. The multivariable situation makes the challenge more difficult. Relevant studies classify multivariate time-series data in three approaches: Raw-data-based (shape-based), feature-based, and model-based [10], [11]. The Raw-data-based method process the raw data directly to evaluate the similarity of two sets of data. The later studies tend to use a set of features or models to represent the raw data, then compare the similarities according to the elements and models [18]. In spite of the existing methods are different, they follow the same general idea, which is to reduce the data dimensions by transforming the raw data into a set of values.

Košmelj [14] adopts the Iterative relocation clustering procedure as a classification method. The author also takes the generalized ward criterion function as the minimization objective to decide the number of clusters and perform optimization. As its contribution, the study [14] proposes the cross-sectional approach to measure the dissimilarity between two sets of multivariate time series data. This approach defines the dissimilarity between trajectories, then proposes a compound interest model to estimate the required time-dependent weights. As a case study, the method is implemented to classify 23 countries into five categories based on their consumption of different types of energy between 1976 to 1982. This method required vast computational resources and storage during the calculation process. In addition, it needs all time-series data to have the same length for comparison.

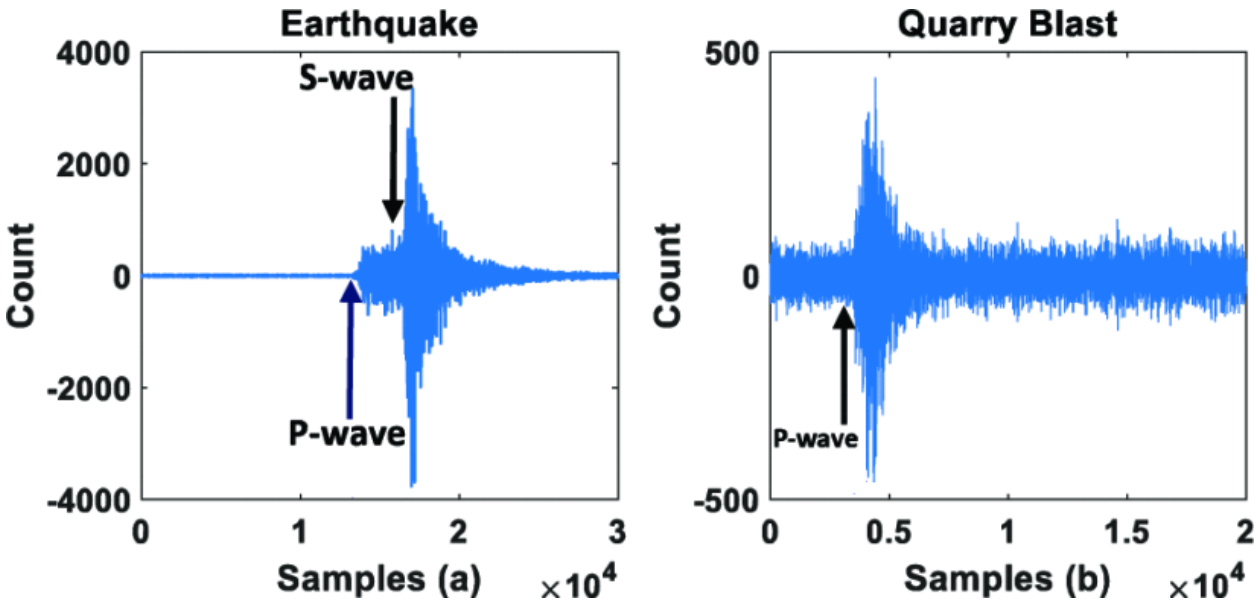

Figure 3.1 [19] Example of earthquake and quarry blast

As shown in Figure 3.1, explosion and earthquake waveform are similar in their time-series records. To distinguish the two events from their multivariate time series data, some features are extracted such as relative amplitude, spectral ratio or relative power components [15]. The article introduces another approach based on the raw data. It applies the Kullback-Leibler (KL) distance and the Chernoff Information Divergence to estimate the difference of sample spectral matrices and group average spectral matrices [15], followed by the K-means and hierarchical classification methods for further classification. However, in the case study, the author has known the data type already and uses the information for validating the proposed classification method.

Like [15] , Shumway [20] also applies the KL discrimination information method to measure the differences between two time-frequency profiles. It adopts the hieratical clustering

classification method after evaluating profiles' similarity. The article claims that the method eliminates the decision about how to extract features. However, the author mentioned that aligning the arrival times is critical for the proposed method.

Biernacki et al. [21] focus on comparing the Bayesian Information Criterion (BIC) and the Integrated Completed Likelihood (ICL) for choosing Gaussian mix model and the number of clusters. The work gives three case studies and claims that the BIC method tends to overestimate the number of clusters when the model does fit the data set well. On the contrary, the ICL method penalizes overlapping clusters. In my opinion, the mentioned BIC's character can be beneficial in some situations. It can detect minor clusters and distinguish partially overlapped clusters.

Ramoni et al. [16] represent multivariate time series data as a set of Markov Chains which describes transition probability of the data. The KL distance is used for measuring the similarity between two sets of Markov Chains, and the similarities are used as a guide for the searching process. The grouping process is to maximize a Bayesian scoring metric of the obtained clustering.

Oates et al. [22] assume that a set of multivariate time series data are generated by Hidden Markov Models (HMM). It first applies the Dynamic Time Warping and the hierarchical classification methods to estimate the number of clusters and the initial clustering. The HMM of each cluster is trained by iteratively moving time series data between clusters until their likelihoods are maximized. To test the effectiveness of the method, the authors use two HMMs

to generate 40 sequences, each sequence with a length of 200 time steps, and then run the classification method on the dataset.

Li et al. [17] also adapt the HMM for the multivariate time series data. The proposed Bayesian HMM clustering method chooses the number of states and clusters by maximizing the BIC. Then, it assigns objects to the corresponding clusters based on its object-to-HMM likelihood. The process finishes when the maximum likelihood achieved. The method is evaluated on both artificially generated data and ecology data using the Partition Misclassification Count metric. However, the testing data is relatively short (i.e., each sample has a length of 56 time steps).

In a study, the multivariate time-series data is assumed to have the Markov property [12]. The data is generated according to a series of unobservable states. The proposed process adopts the HMM method for classification and contains four steps which are determining the number of clusters, the structure for a partition size, the HMM structure and HMM's parameter. To be more specific, it uses the K-means method or depth-first binary division to determine the structure of a give partition size. And the Partition Mutual Information Measure is applied to estimate the number of clusters.

Ferreira et al. [23] propose a new method, namely CPT-M, based on the Principal Component Analysis. The author claims the 24-hour multivariate time series data can be reduced to two principal components while maintaining 95% of the variance of the data. The proposed method implements the PCA similarity matrix, dissimilarity matrix, multidimensional scaling

and date subtractive algorithm iteratively based on the two principal components to classify the studied dataset.

Many methods have been proposed for classifying the multivariate time-series data as shown above. However, none of them are universal. Researchers chose the appropriate similarity measure for the data. And the appropriate measure depends on the nature of data. For example, Li et al.[17] assume their ecology data satisfies the HMM property. Therefore, they find HMM for each set of data and measures the similarity between HMMs. In another study [15], Kakizawa et al. mention the explosion and earthquake's frequency profile are visually different. Therefore, the researcher applies KL distance for measuring the similarity of waveforms' frequency profile.

Among the reviewed studies, researchers have limited ways to verify their proposed method, because they usually do not know the type of each sample before classifying them. In [21], assessment is based on visualization of the classification results. Kakizawa et al. [15] have already known the type of data before the classification process and use the known information to evaluate the classification accuracy. Oates et al. [22] generate testing data using given models. Then, the researcher classifies the generated data and checks if the classification results match the original models. Košmelj, Ferreira and Li apply their proposed method to study cases but they do not evaluate the effectiveness of their proposed classification methods [14], [17], [23].

Even though the reviewed methods are very different, they share the same concept. They tend to reduce the dimensions of the original data, propose that a method to measure the similarity of two samples, and classify samples into corresponding groups. This thesis follows

the concept, to classify PV systems based on their power flow data, nine features are selected to distinguish the PV systems, and the details are shown in the following section.

3.2 Features Extraction

Feature selection is a process of reducing the dimensions of the original data. A site may have multiple energy sources, including PV, grid input, and diesel generator. Sites' operation behaviour may relate to their energy sources. Therefore, the first three features represent the weight of each energy source in a system.

$$F_1 = \frac{E_{tot.PV}}{E_{tot.in}} \tag{3.1}$$

$$F_2 = \frac{E_{tot.Grid.in}}{E_{tot.in}} \tag{3.2}$$

$$F_3 = \frac{E_{tot.Diesel}}{E_{tot.in}} \tag{3.3}$$

Where

$$E_{tot.in} = \sum_t E_{PV}[t] + E_{Grid.in}[t] + E_{Diesel}[t] \tag{3.4}$$

$$E_{tot.Grid.in} = \sum_t E_{Grid.in}[t] \tag{3.5}$$

$$E_{tot.PV} = \sum_t E_{PV}[t] \tag{3.6}$$

$$E_{tot.Diesel} = \sum_t E_{Diesel}[t] \tag{3.7}$$

The fourth feature is the ratio of grid output to the total energy input. The ratio shows the percentage of the energy dumped into the grid. The ratio can tell if the site is grid-tie or stand-alone. The ratio can also distinguish self-consumption sites.

$$F_4 = \frac{E_{tot.grid.out}}{E_{tot.in}} \tag{3.8}$$

where

$$E_{tot.Grid.out} = \sum_t E_{Grid.out}[t] \tag{3.9}$$

The fifth feature is the ratio of battery charged energy over the total system input energy, which shows the battery utilization. Ideally, all input energy into the system should be consumed by the load directly. This may not always true as the charge and discharge process lead to energy loss and battery degradation.

$$F_5 = \frac{E_{tot.Charge}}{E_{tot.in}} \tag{3.10}$$

where

$$E_{tot.Charge} = \sum_t E_{Batt.Charge}[t] \tag{3.11}$$

The sixth feature is the ratio of the system input energy over the load from 10 p.m. to 3 a.m. This feature indicates the site's primary source at night.

$$F_6 = \frac{E_{night.Grid.in}}{E_{night.Load}} \tag{3.12}$$

where

$$E_{night.Grid.in} = \sum_{t\ from\ 10\ p.m.to\ 3\ a.m.} E_{Grid.in}[t] \tag{3.13}$$

$$E_{night.Load} = \sum_{t\ from\ 10\ p.m.to\ 3\ a.m.} E_{Load}[t] \tag{3.14}$$

The last three features are the ratio of discharged energy overload in three periods. The three periods are 1) from 6 a.m. to 10 a.m. 2) from 10 a.m. to 2 p.m. and 3) from 2 p.m. to 6 p.m. Ideally, PV generates energy during the three periods. These features relate to the site's operation mode. Some sites use a battery as a backup energy source. The three values could be low in this

case. Some sites may have high values to maximize the utilization of PV energy. For the peak-shaving sites, they may discharge their battery during a specific period.

$$F_7 = \frac{E_{period1.Batt.Inv}}{E_{period1.Load}} \; F_8 = \frac{E_{period2.Batt.Inv}}{E_{period2.Load}} \tag{3.15}$$

$$F_9 = \frac{E_{period3.Batt.Inv}}{E_{period3.Load}} \tag{3.16}$$

where

$$E_{period1.Batt.Inv} = \sum_{t\ from\ 6\ a.m.\ to\ 10\ a.m.} E_{Batt.Inv}[t] \tag{3.17}$$

$$E_{period2.Batt.Inv} = \sum_{t\ from\ 10\ a.m.\ to\ 2\ p.m.} E_{Batt.Inv}[t] \tag{3.18}$$

$$E_{period3.Batt.Inv} = \sum_{t\ from\ 2\ p.m.\ to\ 6\ p..m.} E_{Batt.Inv}[t] \tag{3.19}$$

3.3 Determining the Number of Clusters

The Gaussian Mixture Model can always get a better representation of distribution by adding more components. In the extreme case, for a database containing N samples, an N-component GMM will perfectly describe the distribution of the database. In this case, the database is classified into N clusters, and each cluster only contains one sample. Such classification does not help us to understand the commons among samples. Therefore, we want to achieve a higher likelihood without adding too many components. In [21], its case studies show the BIC able to distinguish the overlapped clusters. And in [18], it shows that the BIC is more accurate than another common method, Cheeseman-Stutz approximation when the clusters are similar or the number of objects in a cluster is small. Therefore, the Bayesian Information Criterion (BIC) method is applied to determine the number of clusters.

The Bayesian Information Criterion is proposed in 1978 to determine how many components should be included in the mixture. It also helps to determine which covariance parameterization is suitable. Information Criterion is a variation of likelihood. It penalizes the use of parameters. As likelihood goes up with additional parameter used in the mixture model, a penalty term for the number of parameters is subtracted from the likelihood.

$$BIC_{D,N} = 2\log \mathrm{L}_{D,M}(\mathrm{x}|\Psi) - \mathrm{v}\log(\mathrm{M})$$

Where $\Psi = \{\alpha_1, \dots, \alpha_N, \theta_1, \dots, \theta_N\}$ are the parameters of the mixture model for the model D with N components, M is the sample size, and v is the number of parameters, the pair {D, N} which maximizes $BIC_{D,N}$ is selected. In this case, the GMM classify the data into different clusters but remains similarity within each cluster.

Similarly, the BIC also helps us to select the model. As discussed previously, a GMM contains several gaussian distributions, i.e. $\varphi(\mathrm{x}|\theta) \sim N(\mu, \sum)$. Each component represents the distribution of a cluster. These distributions are elliptical in the space, and centred at the mean vector μ and their shape are determined by the covariance matrix $\sum$. The parameterization of the covariance matrix can be obtained from eigen-decomposition in the form of $\sum = \lambda \boldsymbol{DAD}^T$, where λ controls the volume of the ellipsoid, A is a diagonal matrix controlling the shape of the ellipsoid to be more spherical or elliptical, and D is an orthogonal matrix that controls the orientation of the ellipsoid.

In GMM, we have the flexibility to control the parameterization of each covariance matrix. The shape, volume and orientation of each component can be constrained or be variable. Therefore, there are 14 models available with different constrains of parameterization of the covariance matrix [24]. Their features, corresponding model name and the decomposition of the covariance matrix are listed below. Each model has different performance on the same problem. BIC can also be used for selecting the best model for a specific problem by selecting the best pair {D, N} over all possibilities.

Model	$\boldsymbol{\Sigma}_k$	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Variable	Variable

Table 3.1 [24] Parameterizations of the within-group covariance matrix Σk for multidimensional data available in the mclust package, and the corresponding geometric characteristics.

3.4 Site Classification Using GMM

Nine features are selected in Section 3.4. The nine features should be in the range of zero to one. However, there are some outliers in the feature space that take values much greater than

one, although we have processed the data cleaning process as discussed in Chapter 2. It implies undetected errors in the database. To reduce outliers' impact, sites whose features are in the range of [0, 1.5] are kept.

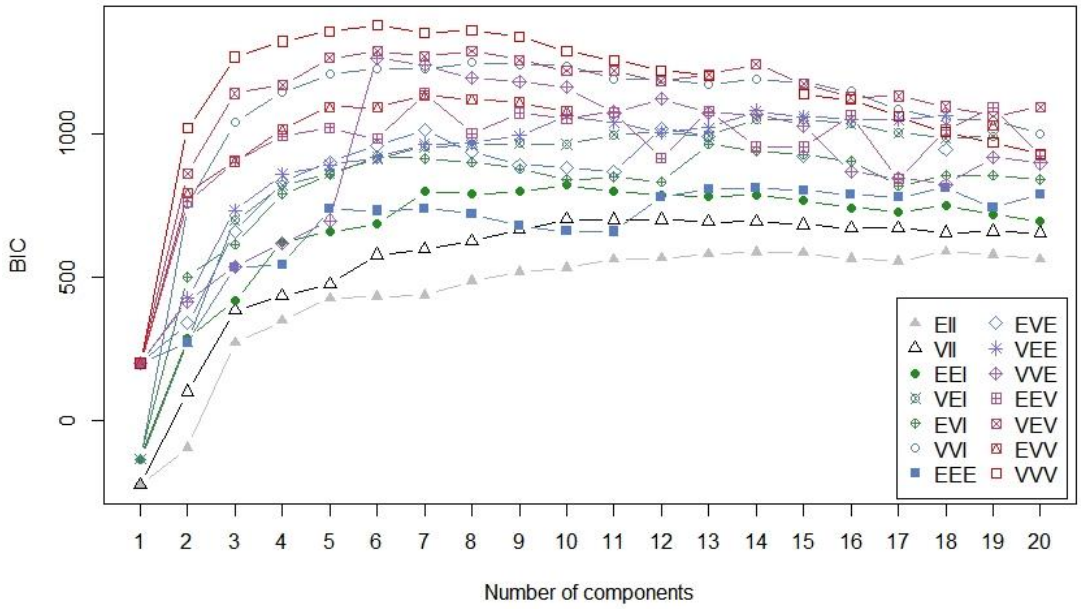

Figure 3.2 BIC plot for models fitted to the proposed features.

To determine the best model and the number of clusters. BIC is calculated for the different number of clusters and different model, as shown in Figure 3.2. According to the BIC method, we expect a dramatic drop once we have reached the best number of clusters. However, it can be seen that the BIC value does not decrease dramatically after reaching the peak point. It implies the systems are not clustered by their nature or the selected features are incomplete. Nevertheless, the volume-variate shape-variate orientation-variate (VVV) model fits the data best when the number of clusters equal to six. The dataset is classified based on the above configurations.

	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5	Cluster #6
# of members	111	79	53	38	28	21
PV (Feature #1)	8.58E-01	0.6058971	5.88E-01	2.72E-01	0.6881349	0.7157959
GEN (Feature #2)	1.41E-01	0.0015475	3.33E-05	2.81E-05	0.0655136	0.1836868
GRIDIN (Feature #3)	1.18E-03	0.3925554	4.12E-01	7.28E-01	0.2463515	0.1005174
GRIDOUT (Feature #4)	1.50E-06	0.1057008	3.48E-01	2.90E-02	0.385891	0.0024344
BATTCHA (Feature #5)	5.91E-01	0.2875284	3.50E-02	2.68E-02	0.2608557	0.5210347
NIGHT (Feature #6)	1.14E-03	0.6976281	1.04E+00	1.03E+00	0.8676186	0.1115871
P1 (Feature #7)	4.28E-01	0.1776547	2.58E-02	7.54E-03	0.6917205	0.3991075
P2 (Feature #8)	7.59E-02	0.0776404	1.10E-02	9.03E-03	0.3574762	0.1399002
P3 (Feature #9)	2.61E-01	0.2179475	2.00E-02	1.91E-02	0.4301292	0.380968

Table 3.2 Means of each cluster in the classification results.

Table 3.2 shows the classification results and the mean value of each cluster. From the mean values, we can get a general understanding of different types of sites in our database. Cluster #1 represents stand-alone sites as there is no grid input. Cluster #3 and #4 are the grid-tie sites which heavily rely on grid input and treat their battery as backup. Their battery stands by for most of the time and does not inject any energy to their systems. Compared to cluster #4, the sites in cluster #3 sell the extra energy to the grid as their grid output is very small. Cluster #6 is also grid-tied, but its features' value close to the cluster #1. Cluster #6 relies on PV generation, as the battery charges more than 50% of the generated energy, and the diesel generator activates sometimes. It implies their grid are not reliable. Compared to the cluster #3, cluster #2 and #5 have more battery utilization. The sites in cluster #5 discharge their battery a lot, which is not preferable. The mean values imply the general characteristics of each type of PV system, but not all details. Each cluster is represented by a multivariate Gaussian distribution, which is defined

by a covariance matrix as shown in appendix table A.1 to A.6. For better illustration purpose, we drew the visualization of those distributions in Appendix B.

The classification results contain six types of systems. From all scatter plots in Appendix B, I can see that most features are widely spread. The selected features are affected by meteorology, load, size and other factors and these factors vary case by case. The major differences are in the following aspects: energy-source priority, grid connection, and if the grid allows PV penetration. The features #7, #8 and #9 do not help in classifying peak-shaving systems. The peak shaving systems may be minorities in the database, or the selected features are not good enough.

The proposed method identified the major PV system types and the outliers in the given database and grouped them into categories. By observation, the six categories are PV/diesel stand-alone systems, grid-tied PV systems, UPS systems that sell energy, UPS systems that do not sell energy, a group of non-conclusive systems and a group of outliers. Among the classification results, the classification of two UPS categories is very successful. Their features concentrate closely in the scatter plots, which indicates a good classification. On the contrary, other cluster's features are scattered on the plots because their power flows are affected by the mentioned factors.

The classification results also indicate some problems. First of all, the meaning of the fifth group is not very clear. My explanation is that the proposed method finds a pattern in the sites of cluster #5, but the pattern does not have any physical meaning. Another problem is that,

in the second group, there are two sub-categories: systems that sell energy and those do not sell energy to the grid. The problem can be eliminated by adding the number of clusters during the classification process. My explanation is that the proposed method tends to pick the major pattern and combines similar minorities into one group. In this case, the sub-category is too small and contains about only 20 systems. Therefore, the sub-category is merged into another cluster.

3.5 Conclusion

The classification for multivariate time-series data is difficult. Previous studies have proposed various method to reduce data dimensionality or to evaluate the similarity between two sets of data. However, there is not a universal solution to solve all classification problems. Regarding the given PV-system data, nine features are proposed to reduce the dimension of the original time-series data while keeping the difference between systems. GMM is also proposed because it is tolerant of the noise of outliers.

Based on the proposed method, six clusters are formed. Four clusters are meaningful. They are PV/diesel stand-alone systems, grid-tied PV systems, UPS systems that are selling energy, a group of non-conclusive system and a group of outliers. As shown in Table 3.3, 281 out of 330 (85%) systems' application types are identified. There are 15% of the system classified as non-conclusive or outliers due to the data quality issue and the randomness of the data's nature.

	Number of Systems in the Cluster	Cluster Type
Cluster#1	111	Stand-Alone PV/Diesel/Battery Hybrid System
Cluster#2	79	Grid-tied PV/Battery Hybrid system
Cluster#3	53	PV/UPS system allowed to inject energy to the grid
Cluster#4	38	PV/UPS system without penetration
Cluster#5	28	Outliers
Cluster#6	21	Non-conclusive

Table 3.3 Classification results and the application of system in each cluster

Chapter 4: Size Optimization of Battery Bank

In chapter 3, four major types of PV systems are found in the given database, in which two types of systems are UPS. The size of the battery in the UPS system is usually decided by the user because it is highly related to the importance of the load and the backup time requirements [25]. Therefore, the battery size in the UPS is not optimized. The size of each component in two of the PV systems, i.e. the stand-alone diesel/PV/battery hybrid system and the grid-tied PV/battery system, will also affect the financial cost. Since the battery's power flow information is available in the database and the battery's degradation is significantly related to its energy cycling [26], this chapter proposes an optimization method for the battery size of the stand-alone diesel/PV/battery hybrid systems (cluster #1 in chapter 3) and the grid-tied PV/battery systems (cluster #2 in chapter 3) using batteries' cycling profile. Moreover, the optimization method is based the following three assumptions: (1) the battery's biweekly cycling profile does not change significantly in any system, (2) the loss of battery capacity is due to energy cycling, and (3) any battery will die in 20 years at a constant rate.

This proposed method determines the size of the battery bank in a given PV system to minimize its weekly battery degradation based on the system biweekly cycling profile and the relation between battery's expected cycle life and its depth of discharge (DOD). To implement this method, a commercial database to calculate the system's cycling profile which describes the amount of energy a battery bank cycles and the number of cycles in two weeks, as well as the expected cycle life vs. DOD from the battery manufacturer, is required. Figure 4.1 is an example of a battery's expected cycle life vs. DOD based on test results. It tells the number of cycles the

battery can withstand at a specific DOD. For example, in figure 4.1, the specific battery can repeat 1150 times cycling at 50% DOD. In other words, each cycling at DOD = 50% costs the battery 1/1150 life.

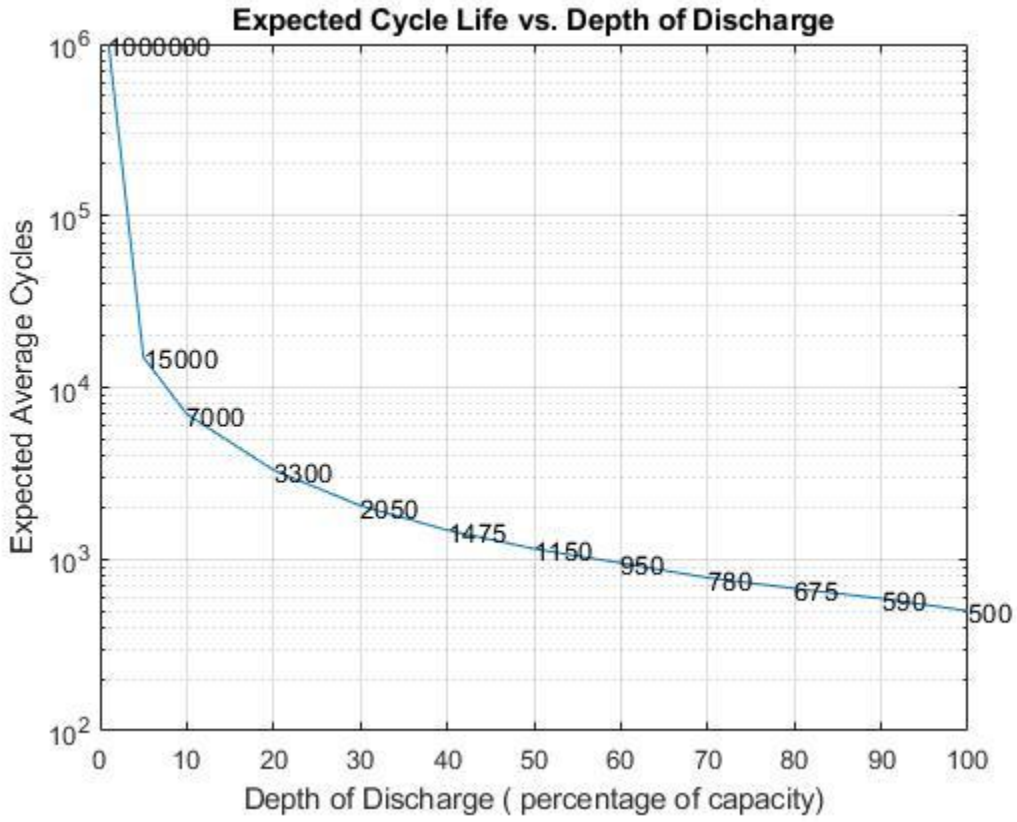

Figure 4.1 An example of expected cycle life vs. DOD from a battery manufactory

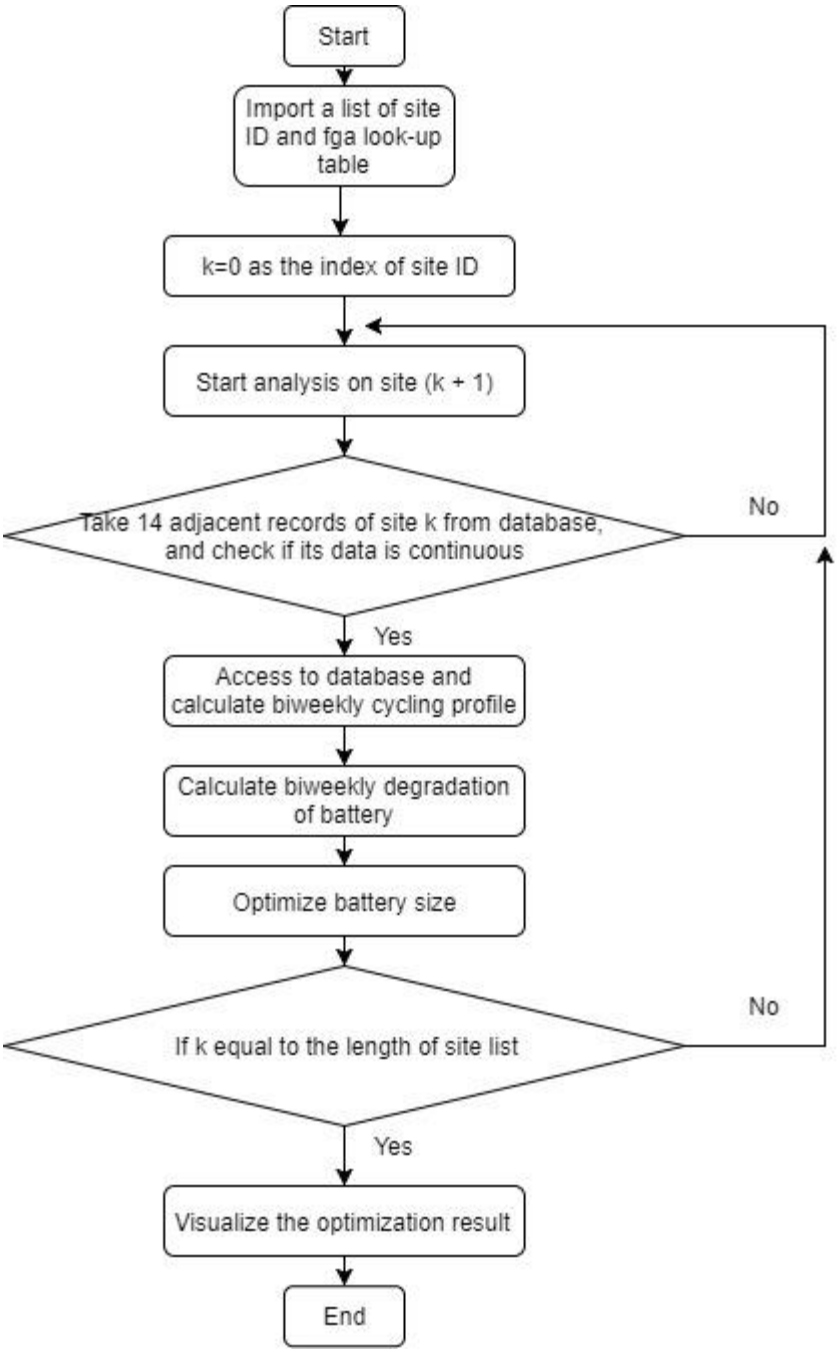

Figure 4.2 The flowchart of the proposed battery size optimization method

The flowchart in figure 4.2 illustrates the procedures of the proposed optimization method. Its main steps are calculating cycling profile, calculating biweekly degradation,

optimizing battery size, and visualizing optimization result. The details of each step are explained in the following sections.

4.1 Calculating Cycling Profile

Chapter 3 has classified the given PV systems into six types. Because only cluster #1 and cluster #2 cycle their battery heavily, this optimization only considers the 190 systems in these two clusters. A new time-series variable is defined to represent the accumulated change of energy in the battery. The new variable's definition is shown below. The energy cycling profile is calculated based on this variable.

$$E_{AccumChange}[t] = \sum_{t'=0}^{t}(E_{BattInv}[t'] - E_{BattChar}[t']) \quad (4.1)$$

Equivalently

$$E_{AccumChange}[t_0] = E_{BattInv}[t_0] - E_{BattChar}[t_0] \quad (4.2)$$

$$E_{AccumChange}[t] = E_{BattInv}[t] - E_{BattChar}[t] + E_{AccumChange}[t - \Delta t] \quad when\ t \neq t_0 \quad (4.3)$$

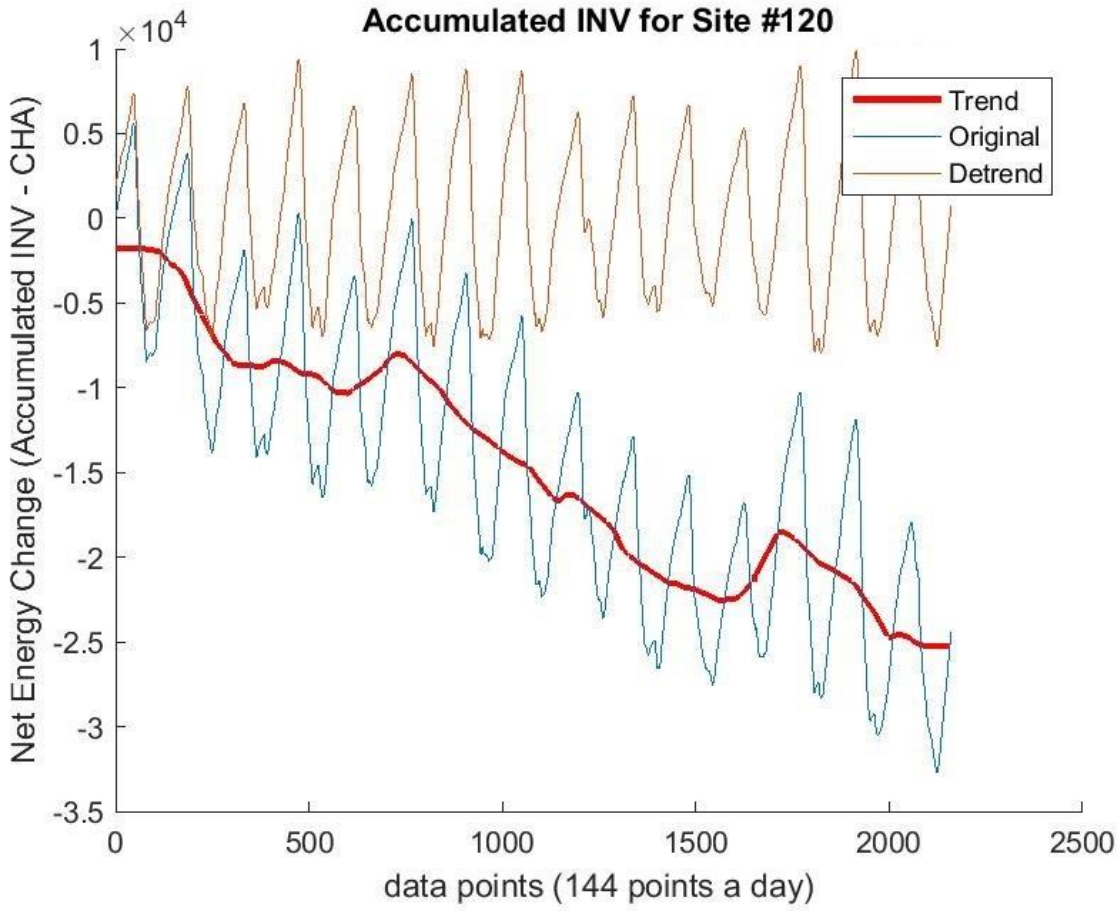

Figure 4.3 Accumulated energy change before and after seasonal adjustment

Figure 4.3 is an example of the accumulated change of energy in the battery over two weeks. The original line is calculated based on the data in the database. The downward trend of the original value is due to the energy conversion lost. Because the charged energy is measured at battery's terminals, it is assumed that the measured charged energy was not converted to chemical energy completely. Therefore, the energy loss should not be treated as energy cycled in the battery. The 13-term moving average method is implemented to remove the seasonal effect. The detrend data will be used for further analysis. From the plot above, the pattern repeats every seven days. Therefore, the biweekly battery cycling profile should contain enough information

which can be used to estimate the battery degradation speed. The rain-flow counting method is applied to calculate the cycling depth and is repeated at each depth [27]. The method converts a spectrum of varying stress into a set of simple stress reversals, which reduces the time-series data to a set of peaks and valleys that are imaged as a pagoda roof. Each peak is imaged as a source that drops water along the roof. The method counts the number of half-cycles when the water flow terminates.

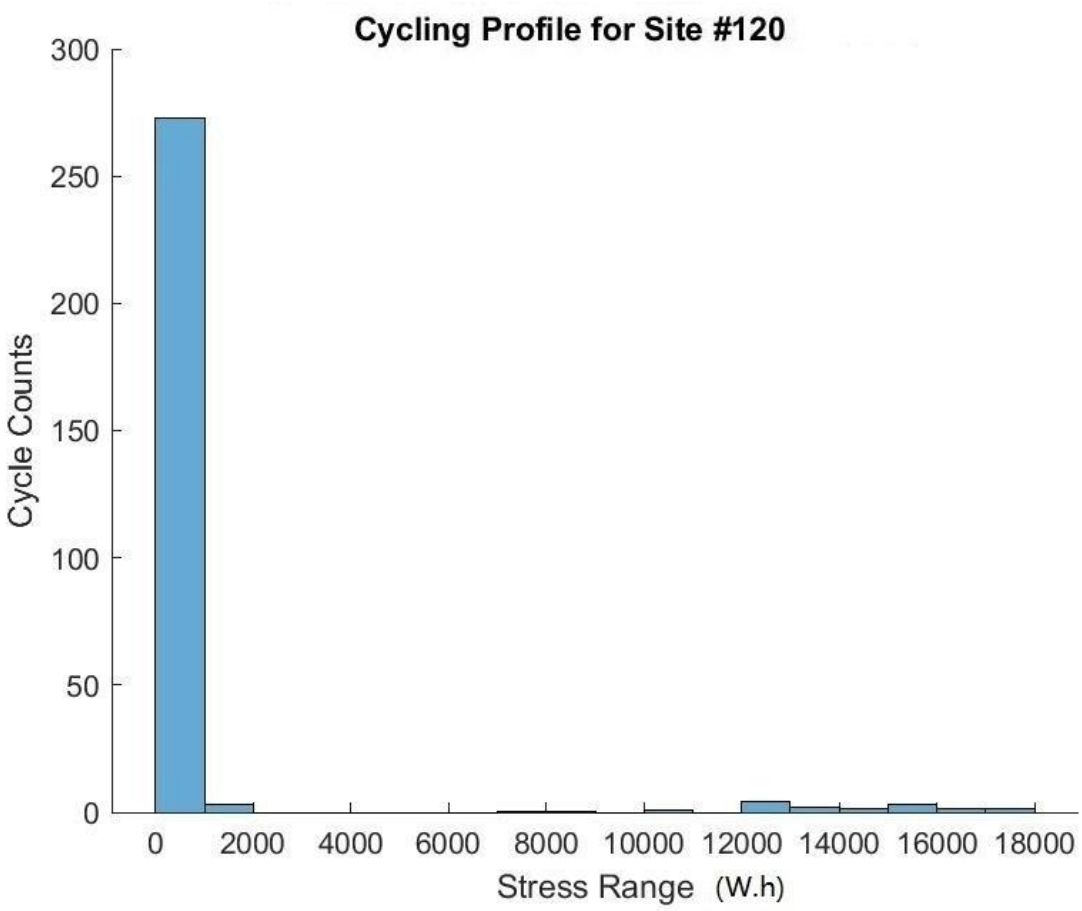

Figure 4.4 An example of battery's year-round cycling profile

Figure 4.4 is an example of a cycling profile that is generated based on the proposed process and year-round power flow data. This battery has about 270 shallow cycling below 2 kW·h and has a few deep cycles around 15 kW·h. If we can get battery bank size in the corresponding system, we will get its cycling DOD profile by using the following formula:

$$DOD = \frac{cycling\ energy}{battery\ size} \tag{4.4}$$

Therefore, once the battery size is known, the x-axis of the cycling profile can be converted to DOD.

4.2 Calculating Biweekly Degradation

To estimate the degradation due to energy cycling, it is necessary to convert the cycling energy to DOD as shown in Equation 4.4. We assume we have the relationship between DOD and expected life, like Figure 4.1. The plot shows a function EC(DOD) that estimates the expected amount of cycles the battery can withstand at specific DOD. Therefore, the degradation due to cycling at DOD can be found as

$$D(DOD) = \frac{1}{EC(DOD)} \tag{4.5}$$

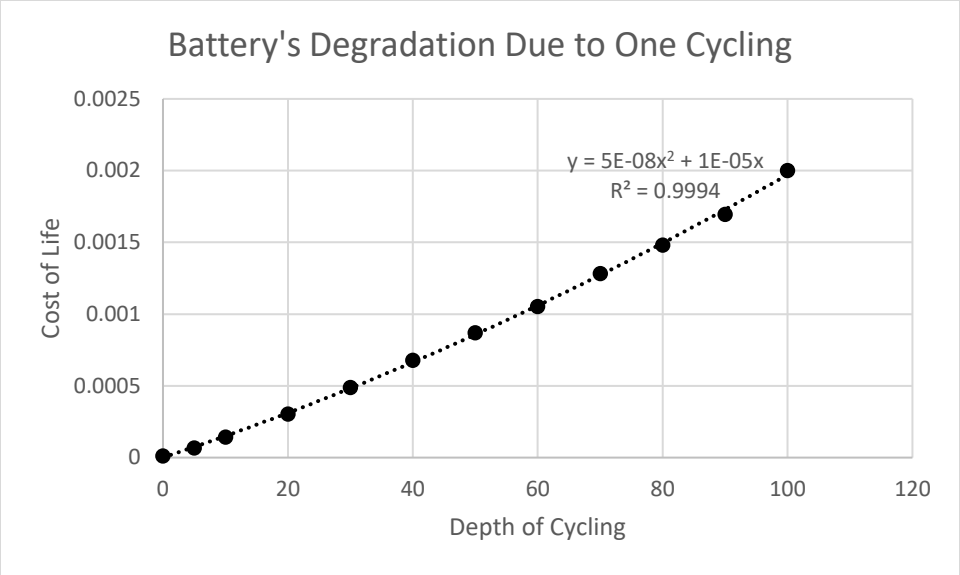

Figure 4.5 Battery's degradation due to one cycling

In Figure 4.5, twelve points are calculated based on the given battery's expected cycle life from Figure 4.1. Then, the function of the expected amount of cycles the battery can withstand at specific DOD can be extrapolated as

$$D(DOD) = \mathrm{a} \times DOD^2 + \mathrm{b} \times DOD \tag{4.6}$$

Where a is 5.1581×10^{-4} and b is 1.456×10^{-3} in this case. In other words, the capacity loss due to one cycling becomes

$$Capacity\ Loss\ (DOD) = Cpacity \times D(DOD) \tag{4.7}$$

If the total n energy cycles are represented by the notation DOD_1, DOD_2, DOD_3……, DOD_n, the bi-weekly capacity loss can be expressed as,

$$Biweekly\ Capacity\ Loss\ Due\ to\ Cycling = \sum_{i=1}^{n} Capacity\ Loss(DOD_i)$$

$$= Capacity\ \times \sum_{i=1}^{n} D(DOD_i) = Capacity\ \times \left(a \sum_{i=1}^{n} DOD_i^2 + b \sum_{i=1}^{n} DOD_i \right)$$

$$= \left(a \sum_{i=1}^{n} \left(\frac{Cycling\ Energy_i}{Capacity} \right)^2 + b \sum_{i=1}^{n} \frac{Cycling\ Energy_i}{Capacity} \right) \times Capacity$$

$$= a \sum_{i=1}^{n} \left(\frac{Cycling\ Energy_i}{Capacity} \right)^2 + b \sum_{i=1}^{n} Cycling\ Energy_i$$

(4.8)

Besides degradation due to cycling, the battery also degrades due to other reasons. According to the previous study, a battery dies in 20 years under the float charge condition [28]. Therefore, I assume the self-degradation rate is 1/20 each year and each year has 52 weeks.

$$Biweekly\ Capacity\ Loss\ Due\ to\ Self\ Digredation = Capacity\ \times \frac{1}{20} \times \frac{1}{26} = \frac{Capacity}{520}$$

Then

$$Biweekly\ Capacity\ Loss$$

$$= Biweekly\ Capacity\ loss\ Due\ to\ Cycling$$

$$+ Biweekly\ Capacity\ Loss\ Due\ to\ Self\ Degradation$$

$$= \left(a \sum_{i=1}^{n} \left(\frac{Cycling\ Energy_i}{Capacity} \right)^2 + b \sum_{i=1}^{n} \frac{Cycling\ Energy_i}{Capacity} \right) \times Capacity$$

$$= a \sum_{i=1}^{n} \frac{Cycling\ Energy_i{}^2}{Capacity} + b \sum_{i=1}^{n} Cycling\ Energy_i + \frac{Capacity}{520}$$

(4.9)

4.3 Optimizing Battery Size

Section 4.2 has formed a quadratic function for biweekly capacity loss with respect to variable Capacity using a biweekly cycling profile. The optimized battery size and the minimized capacity loss can be achieved when the first derivative of the function equals to zero, expressed as

$$Biweekly\ Capacity\ Loss'(Cpacity) = -5.1581 \times 10^{-4} \times \sum_{i=1}^{n} \frac{Cycling\ Energy_i{}^2}{Capacity^2} + \frac{1}{520} = 0$$

Therefore,

$$\text{Optimized Capacity} = \sqrt{0.2682212 \sum_{i=1}^{n} Cycling\ Energy_i{}^2} \qquad (4.10)$$

4.4 Visualizing Optimization Results and Discussion

Figure 4.6 compares the optimized battery size vs. the installed battery size. Each circle represents the battery size in a PV system. The diagonal line represents that the battery size installed same to the optimization result. It can be seen from the diagram that the installed battery size is bigger than the optimized size in most systems. The current battery size in all PV

systems is summarized in table 4.1.

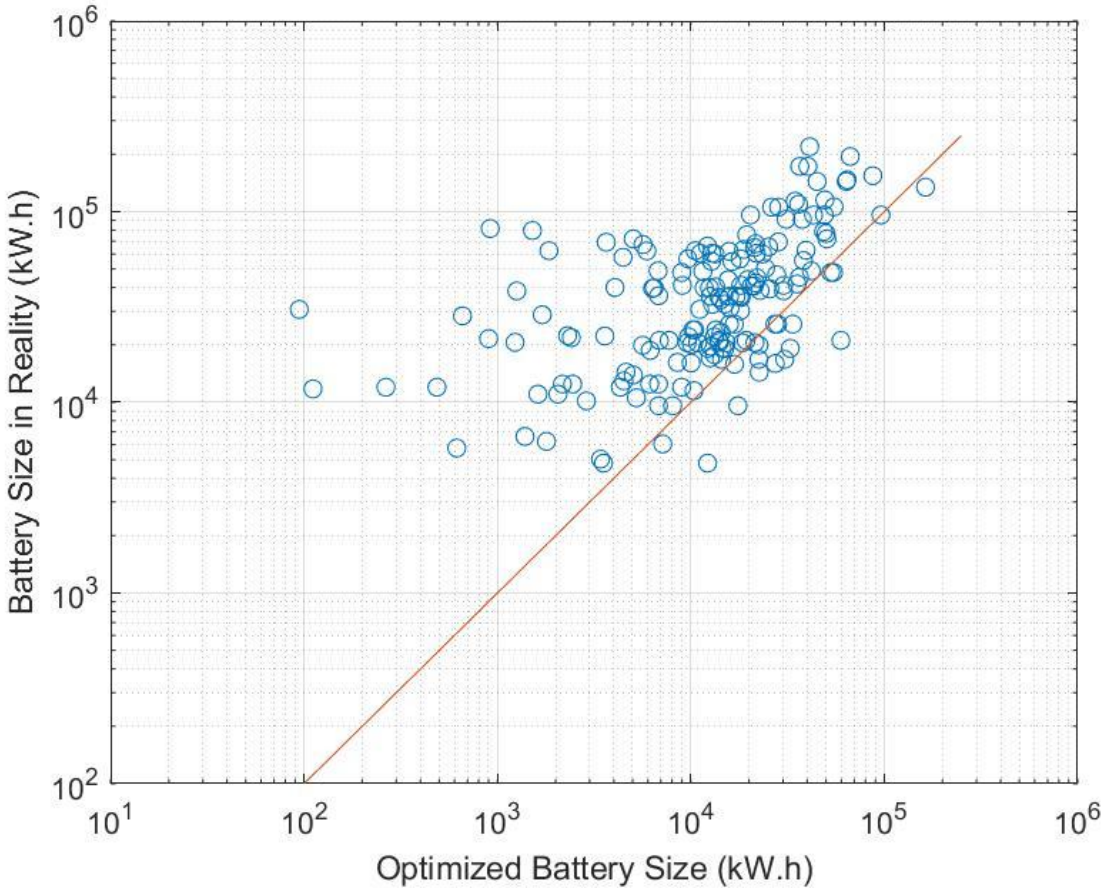

Figure 4.6 Optimized battery size vs. battery size in reality

Regarding the optimum battery size	Number of systems
Undersized systems	19
Invalid systems	20
Oversized more than 10 times	17
Oversized more than 2 times but less than 10 times	84
Oversized less than 2 times	50

Table 4.1 A summary of the installed battery size compared to the corresponding optimal size

Table 4.1 shows that oversize is common in the given PV systems. To understand the effect of battery size on the battery degradation, the power flow data from September 4th, 2017 to September 20th, 2017 in system #26 are used as an example for the following analysis. According to the method proposed in Section 4.2, the biweekly capacity loss due to cycling and the biweekly capacity loss are evaluated with respect to different battery capacities. The evaluation results are plotted in Figure 4.7. As the battery capacity increases from zero, its bi-weekly capacity loss drops. Once it passes the optimum point, the bi-weekly cost increases slowly. In this specific case, when the battery size is 400% of the optimal size, the capacity loss (cost) only increases 20% more than the optimal condition. Therefore, the oversize effect is not significant. The curve without considering self-degradation is monotone decreasing, which is suggesting infinity large battery. So, the speed of self-degradation plays an essential role in the optimization process.

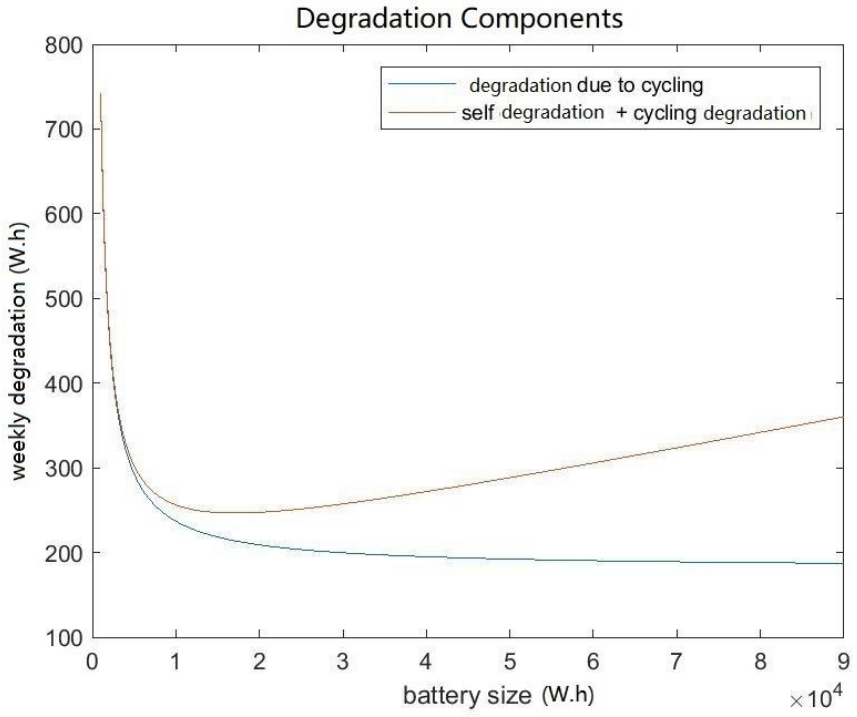

Figure 4.7 Bi-weekly capacity loss functions with respect to the change of battery size

As discussed above, the oversizing does not affect the battery's degradation significantly. To understand the relation between battery size and its degradation, Figure 4.8 is plotted to compare the degradation with respect to the optimum battery size and the degradation with respect to the given battery size. It is also deduced that the oversizing of the battery does not significantly deteriorate the battery's degradation. However, the systems with a small battery tend to size their battery wrongly, which will result in much more capacity losses. Table 4.2 summarizes the degree of degradation among systems.

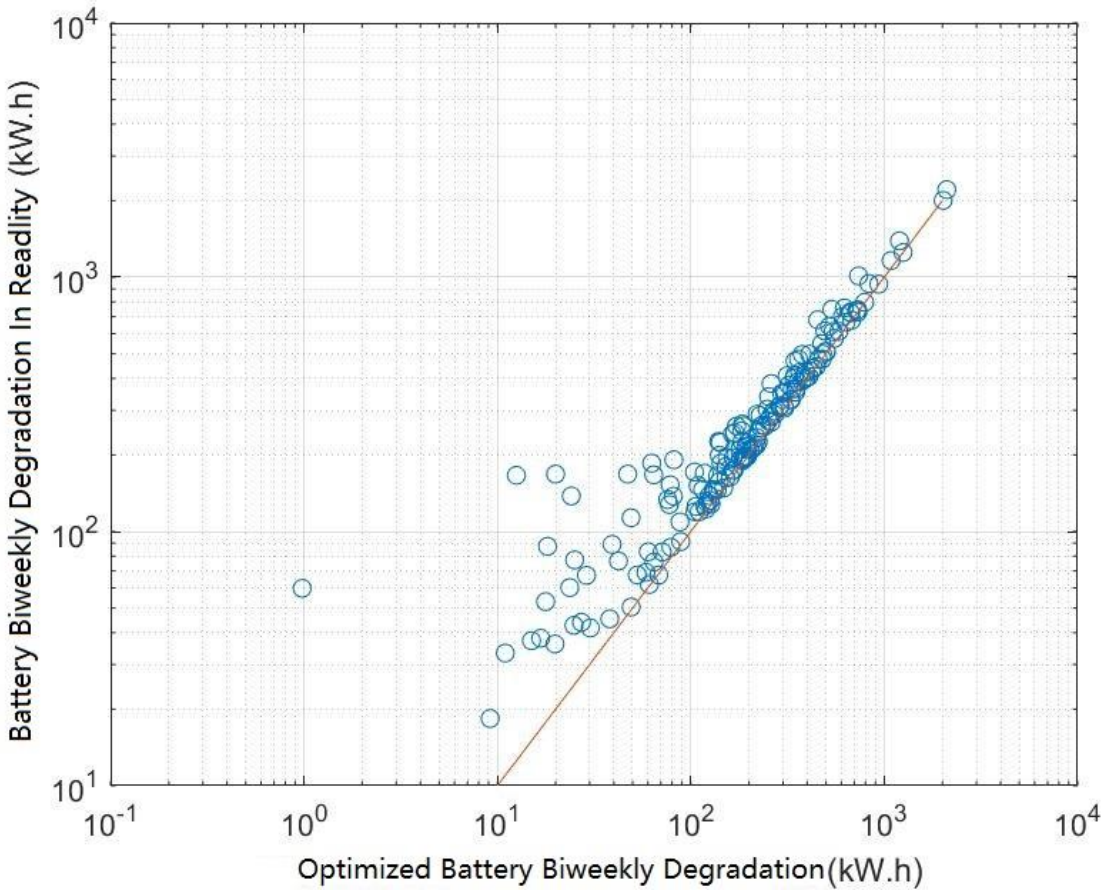

Figure 4.8 Minimized biweekly capacity loss vs. calculated capacity loss based on their battery size

Regarding the minimum battery degradation	Number of systems
Invalid systems	20
Less than 110% of the least degradation	78
110% to 150% of the least degradation	60
150% to 200% of the least degradation	13
More than 200% of the least degradation	17

Table 4.2 A summary of the current battery size with respect to their biweekly degradation of battery

As mentioned earlier, the proposed optimization method assumes the battery's biweekly cycling profile does not change significantly. However, systems' bi-weekly cycling profile can vary over a year. Figure 4.9 is plotted to understand how the change of cycling profile (i.e., each DOD changes) affects the optimal battery size. In the figure, when the profile changes 20%, the optimal battery size varies about 35%.

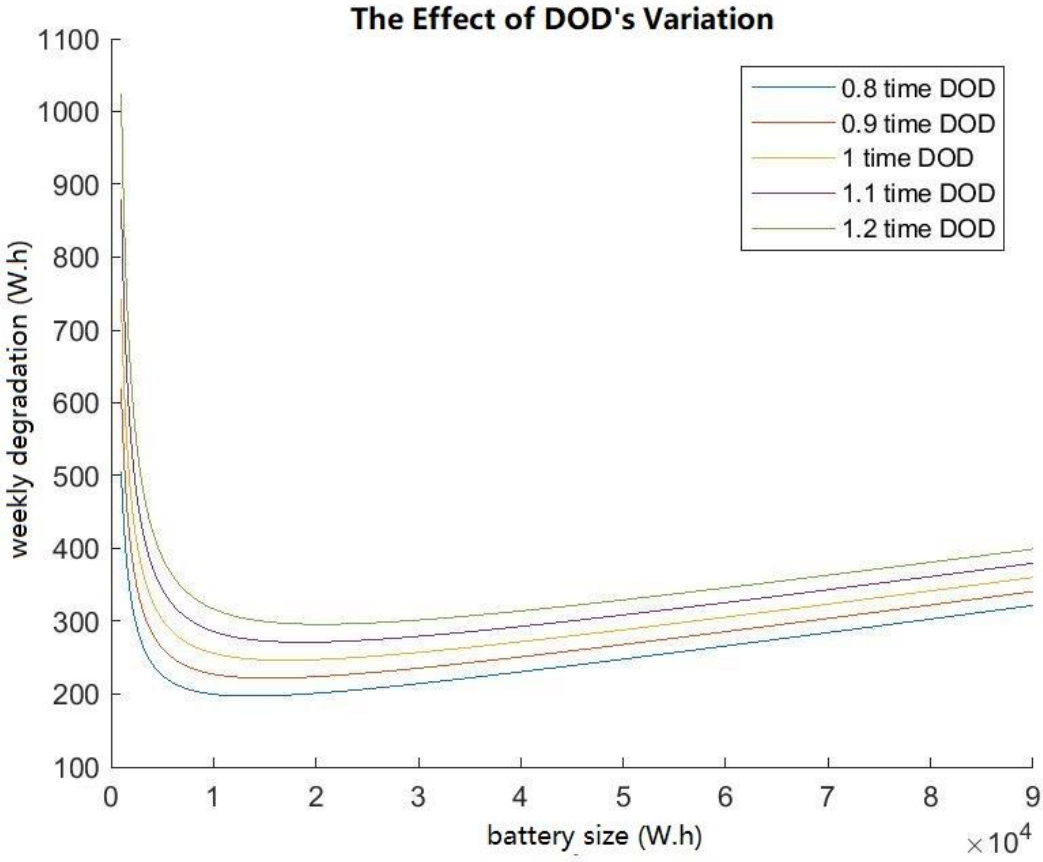

Figure 4.9 The effect of DOD's variation on optimization result

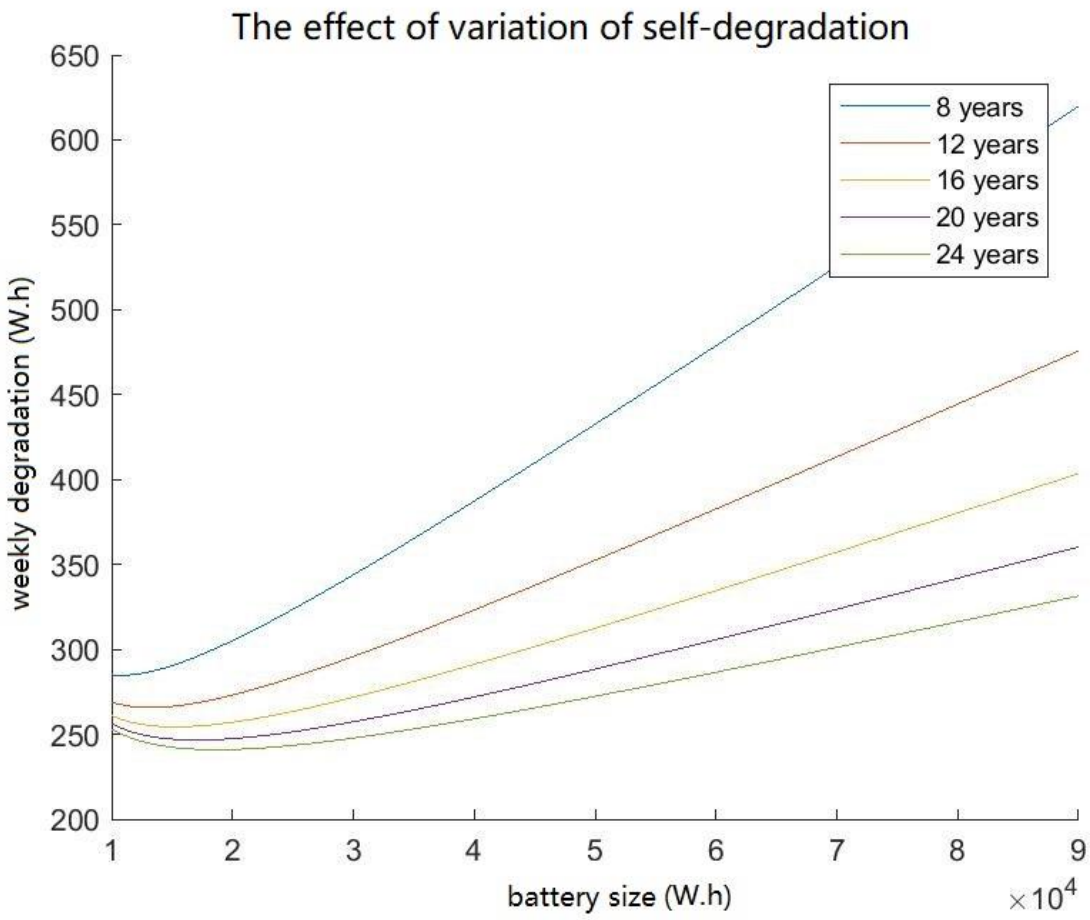

Figure 4.10 The effect of variation of self-degradation on optimization result

As self-degradation decreases (longer life expectancy), the optimal size increases and the right-side slope decreases. Based on the previous study, 10 years to 20 years would be a reasonable estimation [28]. The program was run assuming the life of the batteries being 15 years and 20 years, respectively, and the size of the 15-year batteries is 86% compared to the 20-years.

Overall, in the given database, PV systems tend to oversize their battery bank, but the effect of oversizing battery is not significant. Only a few systems' battery bank are wrongly sized

(i.g., a few battery bank degrade 50% faster than those with optimal battery size). It is worth mentioning that the assumptions made in this chapter restrict the optimization process because the battery degradation is not only caused by self-degradation and cycling. For example, over discharge and overheat are two situations out of the model's consideration [25], [28]. However, the given database does not contain this information.

Chapter 5: Sizing for Multiple System Components

Chapter 4 optimized the battery's size based on their cycling profile to minimize capacity loss due to cycling. However, the PV panel size, diesel generator size can be other important variables. In this case, the battery's cycling profile can be very different, and the previous assumption will no longer be solid. To consider other variables into consideration, this chapter discusses the methods that can size multiple components in a PV system.

In the previous chapters, we identified the major system types in the database. Considering that the systems' applications are different, we may apply a different optimization method. In general, to size the battery capacity for a UPS system, we need to consider the length of reserve time required [29] or interruption avoidance [30]. Regarding the stand-alone system, economic efficiency is usually maximized. And in the grid-tied system's cases, the cost function includes the cost of purchasing electricity and the yield of selling electricity. In some studies, the environmental effect is also considered. In addition, various constraints need to be considered regarding maximum power, maximum current and LLP. In the following case study, a stand-alone PV/diesel/battery hybrid system is optimized using the Artificial Bee Colony (ABC) method based on historical data. The same method can be applied to other types of systems in the database, but the objective function, operation strategy and constraints need to be changed accordingly.

5.1 Review of PV System Sizing Method

In a PV system, we expect the PV array is large enough to satisfy all the load demand. The battery bank capacity should also be large enough to supply energy during the consecutive cloudy days when the PV arrays are not productive. All devices costs include the initial capital cost and the maintenance cost in the latter days. Therefore, to minimize the overall cost, we should size all devices appropriately.

In general, PV system sizing is an optimization problem with respect to various constraints on the parameters including the max area for PV panel, maximum capacity of battery bank or maximum power from the grid. Load demand and radiation information are usually required for solving the sizing problem. In addition, case-dependent aspects should also be considered, such as tariff, system reliability, cost or carbon reduction. Multiple sizing methods have been proposed or reported, such as intuitive method, numerical method, analytical method, evolution algorithm, and neural network. Other methods are based on commercial software, and the implementation is confidential [31], [32], [33].

5.1.1 Intuitive Method

The intuitive method is based on rough estimation, and its process is simple. With the intuitive method, the PV panel should generate enough energy to satisfy the load, and the battery bank should also be able to support the load for a specific period. Previous studies took various factors into count when employing the intuitive method.

Chilundo et al. [29] designs a PV water pump system for a farming house. They estimate system demand based on daily water consumption. A long-term average of radiation is used to size its system. Hereinafter is the brief illustration of this simple case:

In [29], the authors design a PV system for water pumps. It estimates the energy required to supply a water flow by the equation,

$$P_H = Q \times TDH \times \rho \times g, \tag{5.1}$$

where Q is the water flow rate; TDH is called total dynamic head; ρ is the water density, and g is the gravity acceleration.

The PV panel should provide enough power for the water pump in the worst case. Therefore, the required PV panel rating is

$$P_{PV} = \frac{P_H \times G_{REF}}{G_{Globe} \times F_Q \times \eta_Q}, \tag{5.2}$$

where G_{Globe} is the global solar radiance on a horizontal surface; $G_{REF}$is the incident solar radiance under standard testing conditions; F_Q is called quality factor of PV panels, and $\eta_Q$is the efficiency of the water pump.

The study provides additional information for improving the estimation, such as a quality factor of different PV panels, the efficiency of different water pumps, and the average water consumption of different activities. The concept of such a method is simple that generated energy should satisfy the demanded energy in the worst case. However, such a system will waste PV energy in other scenarios. In the proposed equations, many empirical coefficients are

introduced, and the value may be not accurate in a different case. In addition, other studies employed the intuitive method for more complex cases.

Sharma et al. [30] size an off-grid farming house. They suggest a tilt angle and uses average daily load and radiation for sizing panels. They measure the load demand between two consecutive days when the panel cannot provide enough energy for sizing battery. Sidrach-de-Cardona [34] propose the worst-case data should be used for sizing its system, and the system will satisfy its load in all scenario. Bhuiyan et al. [35] size three off-grid systems. They introduce a heuristic parameter, battery autonomy day, describing how long its battery can support its load. However, they assign a value to parameter but does not discuss how to choose the value.

It is worthy to note that the intuitive method is usually for a stand-alone system which has only a battery bank and PV panels. The method tends to introduce heuristic parameters and omit variation and uncertainty of load demand and radiation. The method can satisfy the load demand but fail to discuss how to optimize its size. However, the method is easy to calculate and understand.

5.1.2 Numerical Method

In contrast to the intuitive method, a numerical method concerns the randomness in solar radiation and load demand. It simulates the system's energy flow based on the historical data or simulated data and comes up with an index for comparison. Loss of load probability (LLP) is the popular index in the method as defined as follow:

$$\text{LLP} = \frac{Total\ Unserved\ Energy}{Total\ Demanded\ Energy} \tag{5.3}$$

Tsalides et al.[36] simulate radiation using the formula of Gamier and Ohmura [37]. Based on the simulated radiation and load profile, the LLPs are calculated for the different combination of the tilt angle, battery size and PV array size. The authors select the tilt angle based on three types of the plots: LLP vs. PV array area, PV array area vs. battery capacity, and unserved energy vs. PV array area. However, the paper does not further discuss how to choose the most economical size of the PV array and battery.

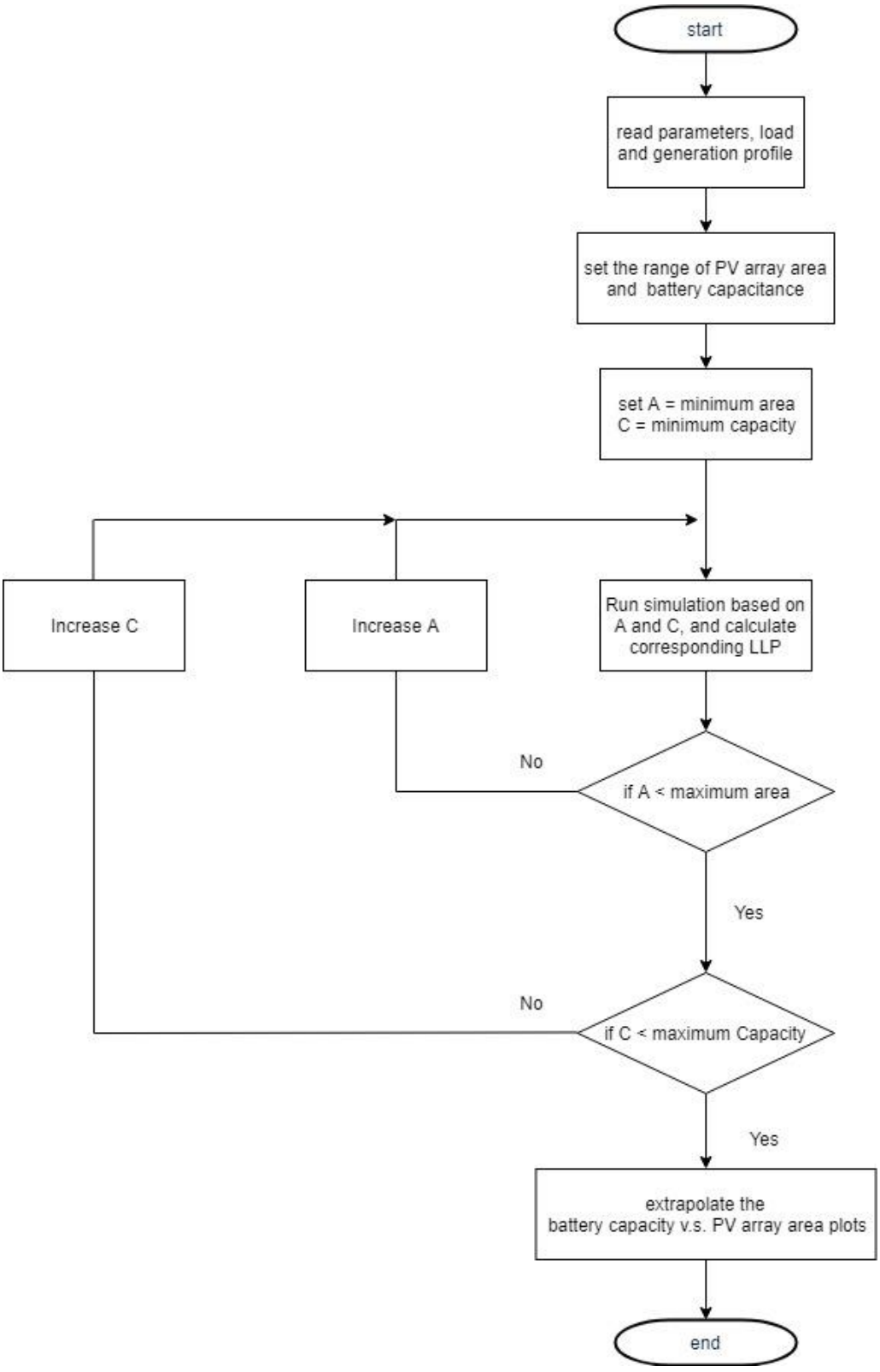

Figure 5.1 A flow chart of the numerical method

As a typical example for the numerical method, Tsalides [36] simulates the solar radiation G(t) using a formula and defines four load profiles for each season. The PV array output is defined as a formula E (A, G(t), θ), where A is the area of PV panel and θ represents all other variables such as temperature, efficiency, solar cell absorptance and other coefficients. Based on the load profile and generation profile, system simulation can be implemented for different combinations of PV array area and battery capacity as shown in Figure 5.1. The system LLP can be derived from the simulation in each scenario. Then the results can be plotted in a graph. Figure 5.2 represents the possible combination of PV array area and battery capacity to make LLP equal to 0. The user can choose the most economical system solution from the curve.

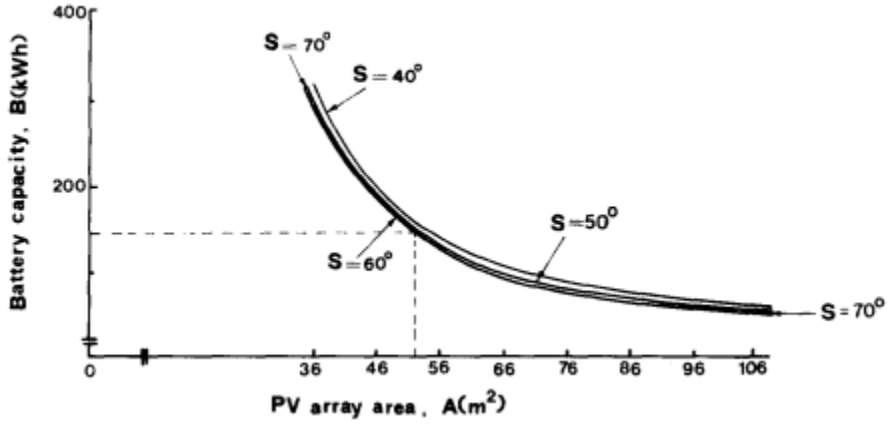

Figure 5.2 [36] Battery capacity vs. PV array area, indicated by least-squares fitting curves, for LLP = 0 and four different tilt angles, S = 40 °, 50 °, 60 ° and 70 °, of south-facing modules

Besides the error in the simulated load profile and generation profile, the method relies on a vast amount of computation. For any pair of battery capacity and PV array area, it needs to simulate for the year-round system status. Theoretically, there is an infinite number of possible

combinations. So, it takes a massive amount of computational resources to plot the accurate curve. The disadvantage can be worse when we process time-series data.

Differently, Shen [38] estimates PV outputs based on average daily radiation, maximum radiation in a day, and radiation hours. It introduces the ratio of PV panel and battery and then proposes a method to achieve the most economical combination of battery size and PV array size. While Kaldellis [39] forces the LLP to be zero and running simulation for different PV size and find the required battery capacity. They formulate the installment cost and finds the cheapest combination. It also notices that, for the specific application, the variation of tilt angle within 45 to 60 degree does not notably affect the overall cost. Similar to [36], Egido [40] also use LLP index and simulation process. However, Egido validates two transposition methods from horizontal to tilted radiation by comparing the sizing results based on historical data and simulated data. They claim that the two transposition methods are not practical. One disadvantage of this method is that it is computationally expensive.

5.1.3 Analytical Method

Analytical methods put the day-to-day radiation variation into consideration. They tend to use a statistical model to describe the randomness and formularize system's behavioir and then optimize the size of the system. However, to apply a statistical model, various assumptions are made that do not always fit reality without justification. Besides, empirical parameters are introduced in formulas. Gordon [41] treats the battery sizing problem as a correlated stochastic problem. Gordon [41] assumes the PV generation in a day has three levels of possibilities: high,

medium and low. It calculates their probability, correlation, and persistence based on historical data. Then, the researcher solves the problem through a formula presented by E. H. Lloyd. With a focus on the long-term energy balance, Markvart et al. [42] design a PV system in London. The author splits year-round data to several climate cycles. Then, corresponding constraints are formed based on the average radiation and the number of consecutive below-average radiation days in each climate cycle. A system sizing formula can be formed that satisfies all constraints. While Demoulias [43] focus on the inverter's sizing for PV systems. It formularizes the inverter's efficiency, the DC power duration curve (PDC) for a specific system, and finally, the inverted power in AC. It finds the optimum inverter size by maximizing inverter's inverted energy. Finally, Arun [44] assumes the PV production in an hour follows a normal distribution. Therefore, it uses mean value, deviation to describe PV inputs. It also introduces a parameter, confidence level, and then convert the probability problem to a deterministic problem. Then, the battery size and PV array size can be determined through simulation. However, its validation shows that the proposed method tends to oversize slightly. As shown above, the analytical methods are very different from each other, and they are usually based on strict assumptions for a specific application. So, the methods are hard to be generalized.

5.1.4 Metaphor-Based Metaheuristic

The metaphor-based metaheuristic is a group of optimization methods which are inspired by nature. It is usually for problems with a large sample size that is impossible to be thoroughly sampled. The method can provide a sufficiently good solution but not guarantee global optimality. Like the Evolutionary Algorithm (EA), it mimics the regeneration, mutation

recombination and selection process in animal's evolution [45]. The algorithm improves the numerical method as mentioned above, because it can solve problems that are constrained, multivariate or multi-objective. The metaphor-based metaheuristics based on much fewer assumptions, so it can be used for a wider range of problems. However, the algorithm contains some parameters need to be configured, and the configuration affects optimization speed and accuracy. To apply the metaphor-based metaheuristic to the PV system sizing problem, we must describe the power functions of each component (i.e., PV panels, battery, interaction with grid and generator), where the power functions are used to describe how the power generated, dissipated or stored. One or more objective functions, thus, can be formed. The metaphor-based metaheuristics can finally optimize the objective function(s) based on historical or simulated load data and meteorological information. For instance, Javadi [46] optimizes for a battery-based wind-turbine/PV system and considers capital, operational and replacement cost. The study tests the artificial bee colony (ABC) method and the particle swarm optimization (PSO) method in one case [46]. It concludes that both methods achieve similar results and the ABC method is faster than the PSO method. While Singh [47] analyzes a grid-tied biomass/PV/diesel hybrid system. It models the grid interaction and discusses how the sale capacity affects the system sizing. Different from the studies as mentioned above, Suchitra [48] applies the adaptive particle swarm optimization method optimizing for two objectives at the same time: minimizing not served energy and minimizing per unit price, rather than assuming LLP, Finally, Hameed [49] add reliability as a constraint of minimizing the system's life cycle cost by open-space particle swarm optimization method. From the papers shown above, we see that many metaphor-based metaheuristics work for the system sizing problem. It enables the analysis of a complex hybrid system.

Similar to the numerical methods, Metaphor-based metaheuristic needs the load profile and generation profile for the simulation process. However, instead of simulating for all possible variables' combinations, it requires at least one objective function and solves for the variables. For example, Javadi [46] focus on minimizing the cost of a PV/ wind turbine/ battery hybrid system. Its objective function represents the system cost with respect to the size of the PV array, battery capacity, and power rating of wind-turbines. In addition, the research [46] models the behavior of PV cells, wind generator and battery for simulation, and it also set the status constrains. Finally, both ABC and PSO methods are applied for solving the combination of variables that minimizes the year-round system cost. Compared with the numerical methods, the metaphor-based metaheuristic is more flexible; the objective can be changed for a different system and optimization purpose. It also allows multiple variables to be optimized all together. The ABC method is chosen for further implementation, and the method is explained in appendix C in detail.

5.1.5 Artificial Neural Network

An artificial neural network (ANN) is a set of interconnected processing units. Each unit takes inputs and passes its processing result as outputs. The connection between the two process units also has its weight. An ANN is good at representing the non-linear relationship and pattern detection. The information or detected pattern is stored in the ANN as interconnection's weights [50]. An ANN does not require formula's format of a pattern. However, the ANN needs a massive amount of data for training, and it is hard to explain the meaning of each weight.

With the ANN method, Khatib et al. [51] use four sites optimization results in [52] as training data for an ANN. It uses longitude, latitude, and LLP as inputs for determining PV/load and battery/load ratios. The fifth site's data is used for testing. The author trained ANN in [51] and achieves a similar result as [52]. Similar to [51], Hontoria et al. [53] train a Multilayer Perception network by backpropagation algorithm based on ten PV systems in Spain. The paper adds the yearly clearness index as an additional input and clams the proposed solution achieve better accuracy than the analytical method. With more site data, Mellit et al. [54] use 36 sites' data for training and introduces genetic algorithm during the training process. The proposed solution is compared with a feed-forward neural network trained by the Levenberg-Marquardt method. The trained networks are tested on four sites. Their proposed method gets a more accurate prediction.

ANN is a model that learns from other studies' optimization results and predict the optimal size for other PV systems. The researchers decide the connections within an ANN, number of layers, training method, inputs and outputs. And these are the main difference among different ANN methods. For example, in [52], it implements a numerical method and finds the optimum PV array area for five PV systems. The optimization results are shown in Figure 5.3. It illustrates the optimal PV array size for five locations to achieve the different LLP requirement. The study [51], then, adopted the results of [52] and designed an ANN. Its ANN model is designed to have four layers and connections are also shown in Figure 5.4. Latitude, longitude and LLP are set to be the input to predict the optimum battery capacity and the PV array size. The information displayed by the curves Johor Baharu, Kuching, Ipoh, and Alor Setar, in Figure

5.3 are used for training ANN model, and the curve of Kuala Lumpur is used for validation. It is compared with the model's prediction to evaluate the model's error. [52] claims that the validation's mean error is only 1.2%.

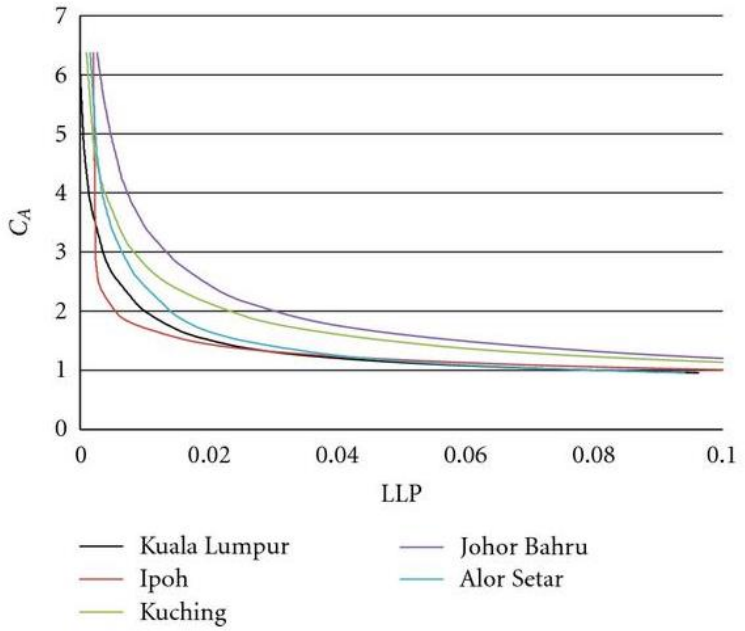

Figure 5.3 The calculated optimum PV array sizes with respect to the constrains of LLP for five PV systems

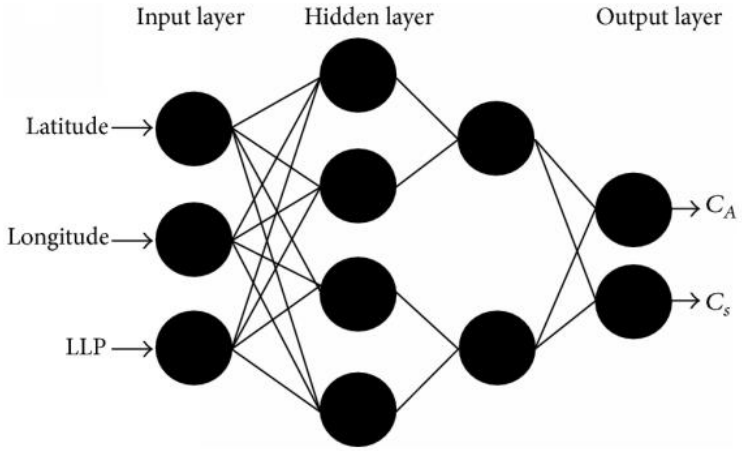

Figure 5.4 A example of ANN proposed by [28] for predicting the optimal PV array size and battery size

The given example successfully predicted other method's result with low error rate. However, it also seems that, to implement the ANN method, other methods should be implemented in advance to obtain the data for training, which seems redundant work if the model is overfit. Another potential problem is that the five locations studied in [52] and [51] are all in Malaysia. So, their climate at the five locations cannot be sufficiently heterogeneous and lose the results generalizability. The two studies did not mention the topology of the five system. Therefore, it will not be an appropriate method in our database, because the systems in the database are all around the world for various applications.

5.2 Problem Formulation

An optimization problem is essentially a mathematical problem looking for the global minimum or maximum of an objective function under certain constraints. In this study, the objective function describes the year-round cost of a PV system, and it must contain all types of costs. Therefore, the maintenance costs, operation costs and capital investment need to be considered. In this chapter, Section 5.2.1 formulates the objective function, Section 5.2.2 models each system component, Section, 5.2.3 discussed the constraints of the problem and the control strategy is defined in Section 5.2.4.

5.2.1 Objective Function

The objective function is the annual cost of a hybrid PV system. The annual cost includes depreciation of assets, maintenance cost, and operation costs such as fuel cost. The system with the lower annual cost is considered as a better system. The function can be expressed as

$$C_{sys} = C_{pv}^{d} + C_{pv}^{m} + C_{bt}^{d} + C_{bt}^{m} + C_{gen}^{d} + C_{gen}^{m} + C_{gen}^{f} + \mathrm{P}, \tag{5.4}$$

where C_{pv}^{d} is the annual depreciation of the PV panel, C_{pv}^{m} is the annual maintenance cost of the PV panel, C_{bt}^{d} is the annual depreciation of the battery, C_{bt}^{m} is the annual maintenance cost of battery, C_{gen}^{d} is the annual depreciation cost of a diesel generator, C_{gen}^{m} is the annual maintenance cost of a diesel generator, C_{gen}^{f} is the annual fuel cost, P is penalty for energy deficit.
The cost analysis is explained in detail in the following sections.

(1) Annual Depreciation

This study makes a simple assumption that the value of the assets depreciates at a constant rate over their lifetime.

$$C_{pv,bt}^{d} = \frac{C_{pu}^{cap} \times \mathrm{N}}{L}, \tag{5.5}$$

where N is sizing variable (P_{PV} or N_{bt}), C_{pu}^{cap} is the capital cost of a unit of N. L is the lifetime of N (year).

Diesel generator's lifetime is in hours. So, its depreciation cost should be rewritten as

$$C_{gen}^{d} = \frac{C_{gen}^{cap} \times P_{gen} \times 365 \times \sum_{t=1}^{s} T(t)}{L_{gen}}, \tag{5.6}$$

where L_{gen} is the rated lifetime of the diesel generator, and T(t) is the generator running time at time step t.

(2) Annual Maintenance Cost

This program assumes that, for PV and battery, maintenance cost is proportional to the device's capital cost, and for a diesel generator, maintenance cost is proportional to its amount of time for the operation. For PV panel and battery, maintenance cost can be expressed as

$$C_{pv,bt}^{m} = C_{pu}^{cap} \times \mathrm{N} \times M_{pv,bt}, \tag{5.7}$$

where $M_{pv,bt}$ is the coefficient for PV panel and battery maintenance cost (/year).

Maintenance cost for diesel generator is

$$C_{gen}^{m} = C_{gen}^{cap} \times P_{gen} \times 365 \times \sum_{t=1}^{s} T(t) \times M_{gen}, \tag{5.8}$$

where M_{gen} is the coefficient for diesel generator maintenance cost (/hour).

(3) Fuel Cost

Energy generated by diesel generator is originally from the chemical energy stored in fossil fuel. The diesel generator converters energy with an efficiency η_{gen}. The amount of fuel can be determined. Fuel cost is

$$C_{gen}^{f} = \frac{C_{gen}^{cap} \times P_{gen} \times 365 \times \sum_{t=1}^{s} T(t) \times P_{diesel}}{H_{diesel} \times \eta_{gen}}, \tag{5.9}$$

where, H_{diesel} is the heat of combustion (kWh/Liter), P_{diesel} is diesel price (/Liter).

(4) Penalty for Energy Deficit

Reliability is crucial for a stand-alone system. The load cannot access any other energy source. So, PV, battery and diesel generator should satisfy all load's need. Energy deficit will penalize the object function. The penalty is expressed as

$$\mathrm{P} = \mathrm{r} \times E_{def}^{2}, \tag{5.9}$$

where r is a coefficient to adjust the penalty. E_{def} is the amount of energy deficit (kW.h).

5.2.2 Component Modeling

The objective function is formulated in Section 5.1.2. The objective function requires the values of the amount of energy deficit, the energy generated by a diesel generator, and the generator running time. These values are obtained based on a time-domain simulation that mimics the power flows in the PV system. To implement the simulation, each component is modelled as follows.

(1) PV panel

The output of a PV panel usually depends on radiation, rated power of the PV panel, temperature and losses due to shade, dirt or temperature. Some models also consider loss due to MPPT. In this study, the PV generation information is collected at MPPT. Therefore, the PV converting rate includes all the factors mentioned above. And the PV output energy at each time step is expressed as

$$E_{PV}(t) = P_{PV} \times G(t), \tag{5.10}$$

where P_{PV} is the size of the PV panel (kW), G is the PV converting rate (h).

(2) Battery

The battery is a passive component in the system. When spared energy generated, battery store energy. Vice versa. Because battery's rating is based on output performance, the program assumes energy losses during charging battery. Therefore, when the battery is charging,

$$SOC_{bt}(t+1) = SOC_{bt}(t) + \frac{E_{pv}(t) - E_{load}(t) - E_{dump}(t)}{N_{bt}} \times \eta, \quad (5.11)$$

where SOC_{bt} is the state of charge of the battery, E_{load} is the energy consumed at load (kWh), E_{dump} is the dumped energy (kWh), η is the battery charging efficiency.

When the battery is discharging,

$$SOC_{bt}(t+1) = SOC_{bt}(t) - \frac{E_{pv}(t) - E_{load}(t) - E_{dump}(t)}{N_{bt}}, \quad (5.12)$$

and the discharged energy during the t^{th} time step can be expressed as

$$E_{bt}(t) = SOC_{bt}(t) - SOC_{bt}(t+1) \quad (5.13)$$

(3) Diesel Generator

A diesel generator is the backup energy source for the system. It operates when PV and battery are insufficient to provide energy. Generated energy in a time step can be expressed as

$$E_{gen}(t) = P_{gen} \times T_{gen}(t), \quad (5.14)$$

where P_{gen} is the rated power for the diesel generator (kW), T_{gen} is the amount of time it operates in a time step (hour).

(4) Invertor

The size of invertor should match the maximum power at the load. The maximum power at load is known. Therefore, the program treated it as a fixed cost and omitted it. This also simplifies the problem by reducing a variable.

5.2.3 Constraints

Power balance constraint, for any time step t, energy injected into the system should be equal to the energy dissipated at loads. The relationship can be represented by

$$E_{pv}(t) + E_{bt}(t) + E_{gen}(t) = E_{dump}(t) + E_{load}(t) - E_{def}(t), \tag{5.14}$$

and the deficit energy is non-negative:

$$E_{def}(t) \geq 0 \tag{5.15}$$

The constraints of the size of the PV panel, battery and diesel generator:

$$P_{PV}, P_{gen}, N_{bt} \geq 0 \tag{5.16}$$

The constraint of the battery capacity:

$$1 \geq SOC_{bt}(t) \geq 0 \tag{5.17}$$

The constraint for diesel generator operation time is:

$$\frac{24}{Total\ amount\ of\ time\ steps\ in\ one\ day} \geq T_{gen}(t) \tag{5.18}$$

5.2.4 Operational Strategy of Simulation

The electricity generated at PV panels have the highest priority because the generation does not cause additional cost. The proposed hybrid system adopts the following strategy:

- If PV panels generate more energy than load demand, the spared energy will be stored in a battery until fully charged.
- If the battery is fully charged, spared energy will be dumped.
- If $E_{pv}(t) + E_{bt}(t) \geq E_{load}(t) \geq E_{pv}(t)$, PV panels and batteries will cooperate and satisfy load demand.
- If $E_{pv}(t) + E_{bt}(t) + E_{gen}(t) \geq E_{load}(t) \geq E_{pv}(t) + E_{bt}(t)$, the diesel generator will start to satisfy load demand. The spare energy will charge the battery.
- If $E_{load}(t) \geq E_{pv}(t) + E_{bt}(t) + E_{gen}(t)$, the system will be unable to satisfy load demand. The energy deficit will be accumulated as a penalty in the objective function.

5.3 Simulation Results and Discussions Case #1

The first case study is based on a set of manipulated data, including a daily PV generation profile and a daily load profile. The data are manipulated based on a real system whose ID is 27 in the given database. Its objective function is formulated according to Section 5.2 to size PV panel, generator and battery bank based on the daily profiles. The Artificial Bee Colony (ABC) method is implemented in this case to find a combination of variables' value that minimize the objective function. Appendix C describes the algorithm of ABC in detail.

	Capital Cost	Maintenance Cost	Fuel Cost	Lifetime
PV panels	3000 $/kW	0.7% of capital cost per year	N/A	20 years
Diesel generator	278 $/kW	0.2% of capital cost per hour	0.9$ / Liter	15000 hours

Battery	150 $/kWh	1% of capital cost per year	N/A	4 years

Table 5.1 Cost parameters for simulation.

Perturbation Coefficient	Population	Max. iteration
1	100	50

Table 5.2 Control parameters for ABC algorithm in case study #1

Diesel heat of combustion	Diesel generator's efficiency	Battery charging efficiency
10 kW/Liter	46%	60%

Table 5.3 Other parameters required by the simulation

The sizing process is implemented in MATLAB. According to the description in Section 5.2.3 and Section 5.2.4, a simulation is run based on the given daily load profile and PV generation profile with a time step size of 10 minutes to simulate the power flows in the system. Many parameters need to be set before running the simulation. Table 5.1 shows the cost parameters for the simulation, Table 5.2 shows the control parameters for the ABC algorithm in the program, and other parameters are listed in Table 5.3.

Figure 5.5 shows that the objective function converges to during the sizing process. The ABC algorithm takes only about 15 iterations to achieve a value at $282.1. After 50 iterations, the objective function outputs $275.4, and the corresponding sizes of the PV panel, battery and generator are determined. Besides the solution shown in Table 5.4, cost analysis is implemented to estimate how money is spent. A sample is shown in Table 5.5.

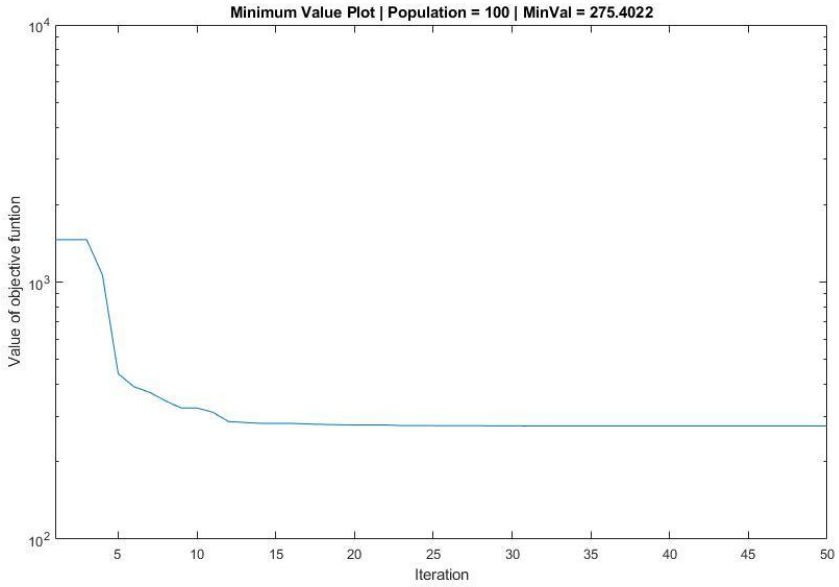

Figure 5.5 Objective function converges to an value after 15 iterations

Size of PV panel	Size of battery	Size of diesel generator	Annual cost
1.5414 kW	0.2993 kWh	0 kW	$275.40

Table 5.4 Sizing result of case study #1

	Annual depreciation cost	Annual maintenance Cost	Fuel Cost
PV panels	$231.21	$32.37	N/A
Diesel generator	0	0	0
Battery	$11.22	$0.45	N/A

Table 5.5 Cost analysis of case study #1

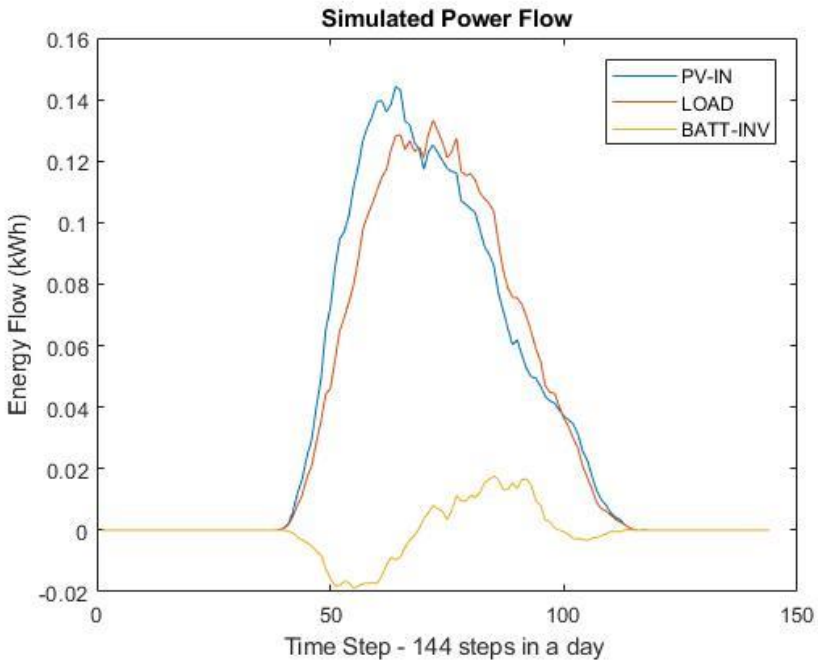

Figure 5.6 Simulated system power flows based on the sizing solution of case study #1

Figure 5.6 shows the simulated system power flows based on the sizing solution listed in Table 5.4. The 'BATT-INV' curve represents the amount of energy inverted from the battery. Its negative value means that the spare solar energy is charged to the battery. The other curves represent the PV generation and load demand. In the figure, the load demands energy and PV panels generate energy during the daytime, but the two time series do not perfectly match. The ABC method sizes the PV panel and battery so that the PV panels generate enough amount of energy to satisfy the demand. Meanwhile, the battery is large enough to buffer the mismatched energy.

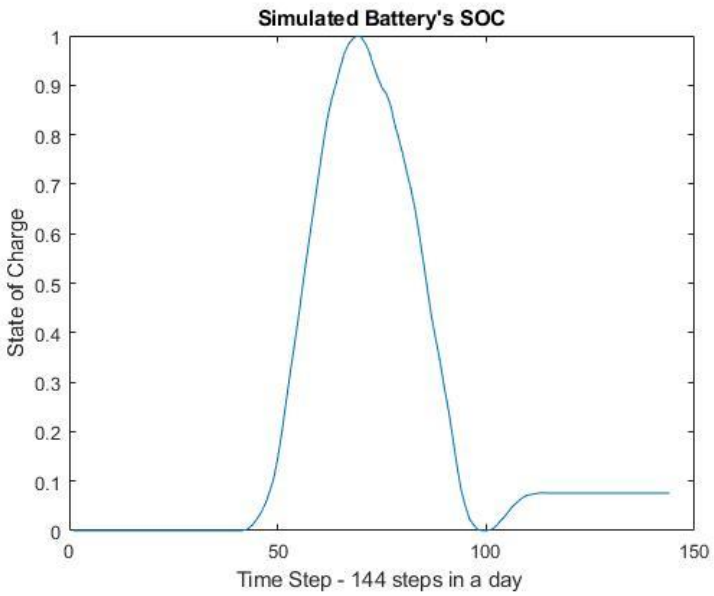

Figure 5.7 Simulated battery usage of the battery in the case study #1

Figure 5.7 shows the simulation result of the estimated battery usage in a day. The battery State of Charge (SoC) is fully utilized from the range of 0 to 1. In other words, the system does not need additional battery capacity. In this case, the simulation result validates the proposed method that the sizing result in Table 5.4 is neither oversized nor undersized. To further validate the proposed method, a more complex case is studied in Section 5.4.

5.4 Simulation Results and Discussions Case #2

Case study #1 is based on manipulated daily load and generation profiles to verify the effectiveness of the sizing method. In the case study #2, the same sizing method is implemented based on real year-round data to size a stand-alone PV system. The system named “red cabin”

(site ID 633) is selected from the given database. The database indicates that the system has installed 10 kW PV panels and a 5 kWh battery bank. However, its recorded peak generation power is only 1.64 kW in the year-round data. Therefore, in this case study, we assume that the system has installed 1.64 kW PV panels. Figure 5.4 only shows seven-day PV generation and load profiles for reader convenience, but the entire year-round data are used for sizing the PV system.

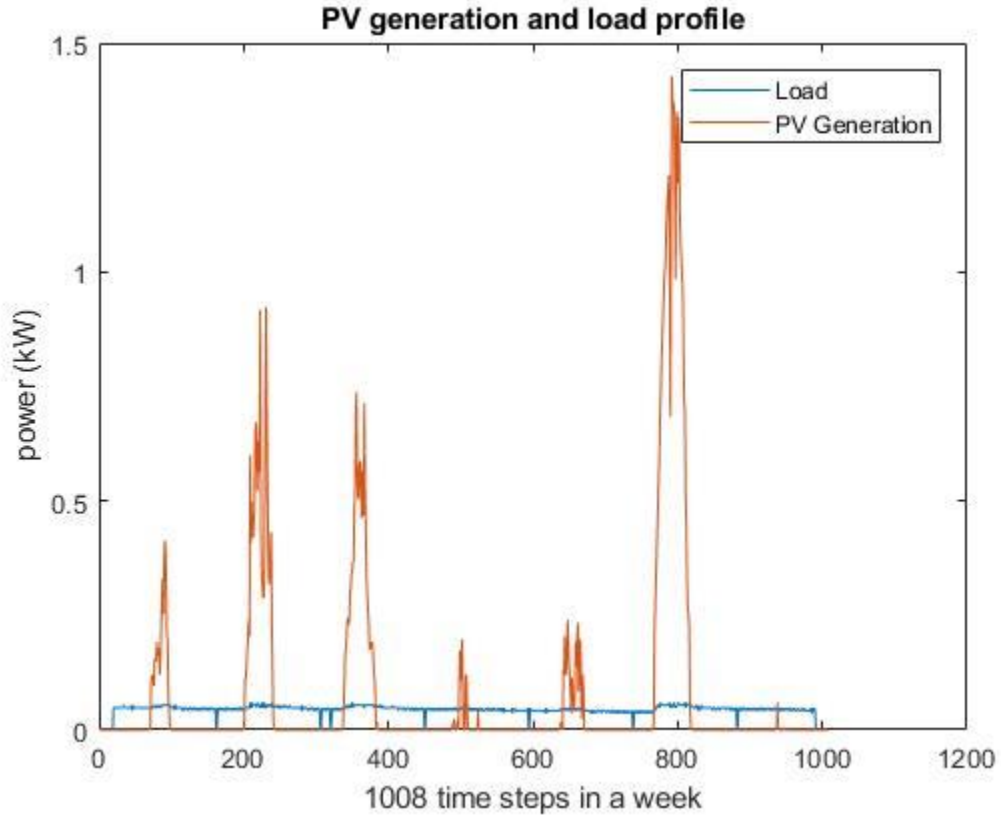

Figure 5.8 Seven-day load and PV generation profile of case study #2

Perturbation Coefficient	Population	Max. iteration
1	300	150

Table 5.6 Control parameters for ABC algorithm in case study #2

Cost parameters used in case study #2 are the same as those used in case study #1 (see Table 5.1 and Table 5.3). Control parameters are changed to adapt to the length of data as shown in Table 5.6. To test whether the method is sensitive for the initial conditions, fifty trials are run. In these fifty trials, PV size, battery size, and generator size are randomly initialized in the intervals of [0,10], [0,50], and [0,50] respectively. All these trials give similar sizing results, which proves that the method is not sensitive to the initial conditions. These fifty sizing results are listed in Appendix D in detail. According to the sizing results, power flows are simulated with a seven-day simulation plotted in Figure 5.9. Notice that the PV generation varies over the seven days. For example, the PV generation reached 2.5 kW on the sixth day, while no energy generated at all on the seventh day. The curve 'BATT-INV' represents the power discharged from the battery, and other time series are for PV generation, load demand, and generator's generation.

Size of PV panel	Size of battery	Size of diesel generator	Annual cost
2.99 kW	21.4 kWh	46.4 W	$1459

Table 5.7 Sizing results of case study #2

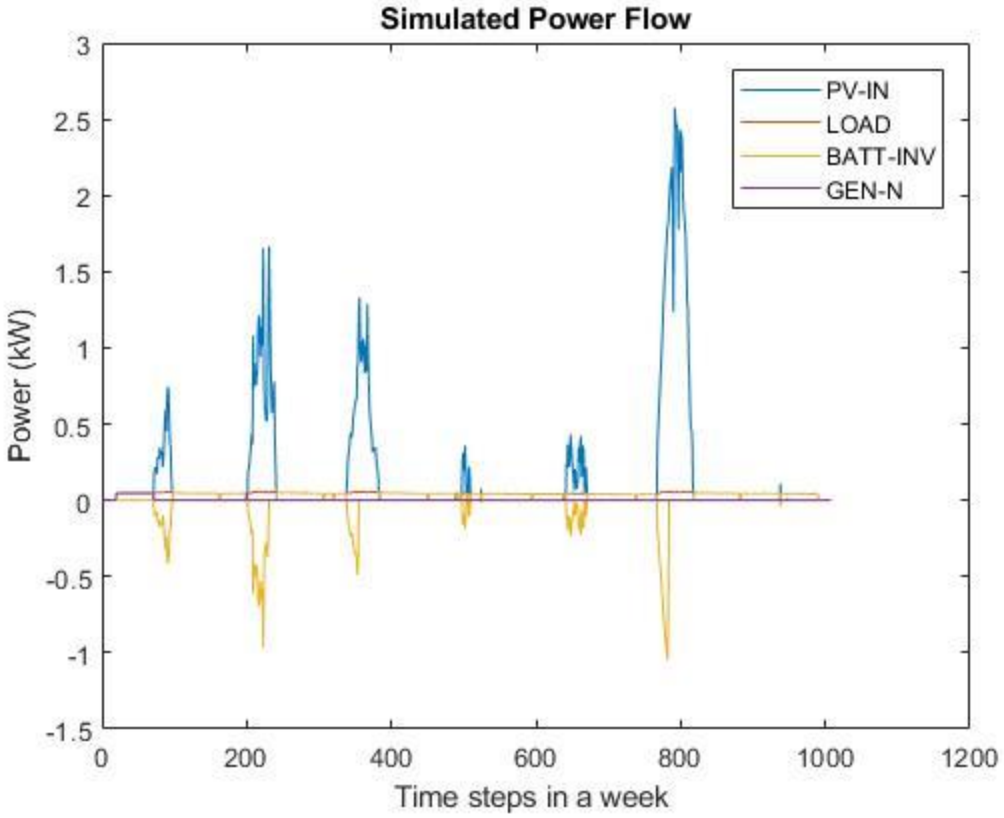

Figure 5.9 Simulated power flows of case study #2 according to the sizing results

To verify the effectiveness of the proposed sizing method, we enumerate 44541 combinations of PV panel sizes, battery sizes, and generator sizes from the intervals [0,10], [0,50], and [0,10] respectively and uniformly. We run a simulation for each of them and calculate the corresponding year-round cost. Among these simulation results, the minimum cost is $1469, which is still higher than the cost given by the proposed sizing method in Table 5.7 at $1459.

	Capital Cost	Maintenance Cost	Fuel Cost	Lifetime
PV panels	4000 $/kW	0.7% of capital cost per year	N/A	20 years
Diesel generator	278 $/kW	0.2% of capital cost per hour	0.6$ / Liter	15000 hours
Battery	150 $/kWh	1% of capital cost per year	N/A	4 years

Table 5.8 Changed cost parameters for simulation.

Diesel heat of combustion	Diesel generator's efficiency	Battery charging efficiency
10 kW/Liter	46%	80%

Table 5.9 Changed other parameters required by the simulation

The effectiveness of the proposed method has been proved, but another test shows that the proposed method is sensitive to the pre-set parameters. To be more specific, the objective function contains many pre-set parameters. When the set of parameters changes, the sizing results will also change. Regarding case study #2, when the parameters change to numbers in Table 5.8 and 5.9, the sizing result will also change. The new sizing result is shown in Table 5.10 and is very different from the previous sizing result in Table 5.7.

Size of PV panel	Size of battery	Size of diesel generator	Annual cost
2.31 kW	21.8 kWh	46.8 W	$1488

Table 5.10 Sizing results of case study #2 with the changed parameters

This chapter discussed methods to select PV panel size, diesel generator size and battery size based on historical power flow data. After reviewing the sizing methods proposed by other researchers, the ABC method is adopted to size a PV system because the ABC method can size multiple variables at the same time for a hybrid PV system. To verify the proposed method, two cases are studied. In the first case, the proposed method can size the PV panel, battery, and generator altogether based on a given PV generation and load profiles. In addition, the sizing result matches intuition. In the second case, the proposed method is tested on real year-round data. Enumeration method compares the sizing result of the proposed method with all 44541 size combinations and shows that the cost of the proposed method is the lowest among all test cases. However, this method is proved to be quite sensitive to pre-set parameters. Therefore, to better implement the method to the given database, more accurate data need to be collected for better results.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

This thesis is based on historical power flow data from the commercial database provided. We address the data quality issues then propose a validation rule and a correction method. The GMM classification method is implemented to detect the systems whose battery monitors are installed incorrectly, the data are then corrected accordingly. The thesis also propose a set of features to identify the type of PV systems using the GMM method. Knowing the type of the PV system is crucial for system sizing because the sizing process needs that information to form the corresponding objective function. In addition, some systems cannot be optimized based on the generation and load profiles. Like UPS, the battery size depends on the user's preference. Finally, two sizing methods are proposed. One is to optimize battery size based on its cycling profile. The other is to size multiple components in a PV system using the Artificial Bee Colony method.

The issues regarding the data quality in the database are addressed in Section 2.1. The law of conservation of energy is applied to verify the data. In addition, the systems are removed from further analysis if there are too many records missing, marked as a dummy site or had too short a history. As a result, only 290 out of the 4000 (5.8%) systems provided are selected for further analysis. Evidently the database has serious data quality issues. Many of the selected 290 systems have their battery monitors connected incorrectly. The thesis proposes three features and implements the GMM method to detect the systems. Then we correct their power flow data.

It is necessary to know the type of the PV system in order for us to size the system correctly. In Chapter 3, nine features are chosen to reduce the dimensionality of the power flow data while keeping each site's characteristics. The GMM method is implemented to classify the 290 PV systems. Six clusters are classified (see Table 3.3). Four main clusters are clearly identified, they make up 83% of the valid data analyzed in this thesis. The four clusters represent stand-alone PV/diesel/battery/hybrid systems, grid-tied PV/battery hybrid systems, PV/UPS systems that inject energy to the grid, and PV/UPS systems without grid penetration. The classification process also provids a cluster of outliers and a cluster of unknown operation. The UPS systems cannot be optimized in this thesis because their battery sizes depend on their design requirements which could be specific to each application, we are not provided with this type of information.

Chapter 4 proposes an optimization method for sizing batteries for stand-alone PV/diesel/battery/hybrid systems and grid-tied PV/battery hybrid systems. It calculates the cycling profile for each system based on their battery charging power flow and battery discharge power flow. Then it derives an optimum battery size that minimized battery degradation due to cycling. As shown in Table 4.1, we can detect the whether the existing battery is too big or too small for each system and suggest the optimal size to the customers. Oversizing of batteries is very common in the given database, though this is not ideal, the model described in Chapter 4 shows that oversizing the battery will not increase the cost as much as undersizing it.

Chapter 5 explores the possibility of sizing several system components simultaneously, including PV panels, generators and batteries. We propose a sizing method based on the ABC method for the stand-alone PV/diesel/battery/hybrid systems. Two case studies done in this thesis verified the effectiveness of the proposed sizing method. However, case study #2 shows that the proposed method is sensitive to other parameters which are not available in the given database. Due to the missing information, we cannot provide a conclusion for this specific set of data through this method. In comparison, Chapter 4 only optimizes the battery size based on its cycling profile to minimize the capacity loss due to cycling.

6.2 Future Work

First the database has some quality issues. From the validation result we can conclude that only a small portion of the data can be justified. Based on observations, missing data and scaled measurements are common in the database. On top of the wrong measurements, system modifications, additional unmonitored devices or multifunctional devices can also cause data quality issues. In many situations, we cannot figure out what causes missing or invalid data by simply reviewing the database. A feasible method is to add a verification procedure in the ComBox or SCP, which can warn the user once their measurements cannot be justified. The company can also provide the user with the system topology templets and let the user update which templet he or she adopts. In this way we are able to verify if their system is correctly installed.

The manufacturers of MPPTs and inverters should inform the customer of the equipment's lifespans. This information is required for accurate cost estimation and system sizing process. The battery bank's life span is highly dependent on how it's used in the system; the ambient temperature, maintenance and battery type are also important factors. Therefore, the battery bank's degradation model should be studied further. Battery manufacturers usually provide the battery life at specific depths of discharge. Current studies tend to choose a formula based on batteries' lab performance. These tests are done under strict conditions at set temperatures and depths of each cycle. However, in real-world applications many other factors contribute to the degradation of batteries, making the given formulae too ideal.

Accurate historical data can help us provide more realistic models of battery degradation. We can do this by using data-based modelling because of its advantage in solving non-linear problems. At this stage, the installed battery monitors are monitoring some very useful data such as SoC, current, voltage, estimated time for discharging, but none of this information is stored in the database. Some software modifications may enable data collection. The change in SoC and power usage can estimate the battery bank's capacity. Once data are collected on these parameters it is possible to study how the temperature, usage, current, and maintenance would affect the degradation and therefore the sizing of the battery. Typically a voltage lookup table is used to roughly estimate a battery's SoC. Using additional measurement devices such as an internal resistance measurement or a SoH measurement would provide a more accurate estimates. We should emphasize that it's crucial to have information about the battery's condition (eg. SoH, remaining life span, remaining capacity) before studies of battery degradation can be conducted.

Chapter 5 proposed the ABC method to size a PV system's components. This method requires many additional parameters. I made reasonable assumptions in the two case studies, they are listed in Table 5.1, 5.3, 5.8 and 5.9. As discussed in case study #2, the sizing result is sensitive to these parameters. Without knowing the accurate parameters of each PV system, I cannot draw a conclusion of whether a device is oversized or not. For further analysis, additional information such as the distribution line capacity, buying price, selling price, grid connection and cost of outage is worth obtaining.

References

[1] M. Eltawil, Z. Zhao, "Grid-connected photovoltaic power systems: technical and potential problems: a review," *Renewable Sustainable Energy,* vol. 14, pp. 119-29, 2010.

[2] P. Sharma, H. Bojja, P. Yemula, "Techno-Economic Analysis of Off-Grid Rooftop Solar PV System," in *IEEE 6th International Conference on Power Systems (ICPS)*, 4-6 March 2016.

[3] M. Schellander, T. Khatib, W. Elmenreich, D. Egarter, "Techno-economical assessment of grid-connected photovoltaic power systems productivity in summer season in Klagenfurt Austria," in *IEEE International Conference Power & Energy (PECON)*, 2014.

[4] G. Zubi, R. Dufo-López, G. Pasaoglu, N. Pardo, "Techno-economic assessment of an off-grid PV system for developing regions to provide electricity for basic domestic needs: A 2020–2040 scenario," *Applied Energy,* vol. 176, pp. 309-319, 2016.

[5] D. O. Akinyele, R. K. Rayudu, N. K. C. Nair, "Development of photovoltaic power plant for remote residential applications: The sociotechnical and economic perspectives," *Applied Energy*, vol. 155, pp. 131-149, 2015.

[6] B. Wichert, "PV-Diesel hybrid energy systems for remote area power generation - a review of current practice and future developments," *Renewable and Sustainable Energy Reviews,* vol. 1, pp. 209-28, 1997.

[7] R. Kumar, A. Agarwala, "Renewable energy technology diffusion model for techno-economics feasibility," *Renewable and Sustainable Energy Reviews*, vol. 54, pp. 1515-1524, 2016.

[8] N. Li, K. Hedman, "Economic assessment of energy storage in systems with high levels of renewable resources," *IEEE Trans. Sustain. Energy,* vol. 6, pp. 1103-1111, 2015.

[9] A. Tomar and S. Mishra, "PV energy benefit estimation formulation for PV water pumping system," in *4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, Mathura, 2017.

[10] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition,* vol. 38, pp. 1857-1874, 2005.

[11] S. Aghabozorgi, A. Seyed Shirkhorshidi and T. Ying Wah, "Time-series clustering – A decade review," *Information Systems,* vol. 53, pp. 16-38, 2015.

[12] C. Li and G. Biswas, "Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification," in *Advances in Intelligent Data Analysis: Third International Symposium*, Amsterdam, The Netherlands, 1999.

[13] H. He and Y. Tan, "Unsupervised Classification of Multivariate Time Series Using VPCA and Fuzzy Clustering with Spatial Weighted Matrix Distance," *IEEE Transactions on Cybernetics,* pp. 1-10, 2018.

[14] K. Košmelj and V. Batagelj, "Cross-sectional approach for clustering time varying data," *Journal of Classification,* vol. 7, pp. 99-109, 1990.

[15] Y. Kakizawa, R. Shumway and M. Taniguchi, "Discrimination and Clustering for Multivariate Time Series," *Journal of the American Statistical Association,* vol. 93, pp. 328-340, 1998.

[16] M. Ramoni, P. Sebastiani and P. Cohen, "Multivariate Clustering by Dynamics," in *Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence*, Austin, Texas, USA, 2000.

[17] C. Li, G. Biswas, M. Dale and P. Dale, "Building Models of Ecological Dynamics Using HMM Based Temporal Data Clustering — A Preliminary Study," in *Advances in Intelligent Data Analysis*, Cascais, Portugal, 2001.

[18] C. Li, "A Bayesian Approach to Temporal Data Clustering using Hidden Markov Models Methodology," Department of Computer Science, Vanderbilt University, Nashville Tennessee, December 2000.

[19] O. M. Saad, K. Inoue, A. Shalaby, L. Sarny and M. S. Sayed, "Autoencoder based Features Extraction for Automatic Classification of Earthquakes and Explosions," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, Singapore, 2018.

[20] R. Shumway, "Time-frequency clustering and discriminant analysis," *Statistics & Probability Letters*, vol. 63, pp. 307-314, 2003.

[21] C. Biernacki, G. Celeux and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 719-725, 2000.

[22] T. Oates, L. Firoiu and P. Cohen, IJCAI-99 Workshop on Neural, Symbolic, and Reinforcement Learning Models for Sequence Learning, 1999.

[23] A. Ferreira, C. Cavalcante, C. Fontes and J. Marambio, "A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector," *International Journal of Electrical Power & Energy Systems,* vol. 53, pp. 824-831, 2013.

[24] L. Scrucca, M. Fop, T. Murphy and A. Raftery, "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models," *The R Journal,* vol. 8, p. 289, 2016.

[25] I. Alaperä, S. Honkapuro, V. Tikka and J. Paananen, "Dual-purposing UPS batteries for energy storage functions: A business case analysis," *Energy Procedia,* vol. 158, pp. 5061-5066, 2019.

[26] P. Ruetschi, "Aging mechanisms and service life of lead–acid batteries," *Journal of Power Sources,* vol. 127, pp. 33-44, 2004.

[27] Q. Deng and H. Yuan, "The Algorithm for Graphic Method of Rain-Flow Counting in Programming," *Advanced Materials Research,* Vols. 225-226, pp. 1157-1161, 2011.

[28] R. Dufo-López, J. Lujano-Rojas and J. Bernal-Agustín, "Comparison of different lead–acid battery lifetime prediction models for use in simulation of stand-alone photovoltaic systems," *Applied Energy,* vol. 115, pp. 242-253, 2014.

[29] R. Chilundo, U. Mahanjane and D. Neves, "Design and Performance of Photovoltaic Water Pumping Systems: Comprehensive Review towards a Renewable Strategy for Mozambique," *Journal of Power and Energy Engineering,* vol. 6, pp. 32-63, 2018.

[30] V. Sharma, A. Colangelo and G. Spagna, "Photovoltaic technology: Basic concepts, sizing of a stand alone photovoltaic system for domestic applications and preliminary economic analysis," *Energy Conversion and Management,* vol. 36, pp. 161-174, 1995.

[31] M. Morad, M. Nayel, A. A. Elbaset and A. I. A. Galal, "Sizing and Analysis of Grid-Connected Microgrid System for Assiut University Using HOMER Software," in *2018 Twentieth International Middle East Power Systems Conference (MEPCON)*, Cairo, Egypt, 2018.

[32] T. Khatib, A. Mohamed and K. Sopian, "A Software Tool for Optimal Sizing of PV Systems in Malaysia," *Modelling and Simulation in Engineering,* pp. 1-11, 2012.

[33] S. Sinha and S. Chandel, "Review of software tools for hybrid renewable energy systems," *Renewable and Sustainable Energy Reviews,* vol. 32, pp. 192-205, 2014.

[34] M. Sidrach-de-Cardona and L. Mora López, "A simple model for sizing stand alone photovoltaic systems," *Solar Energy Materials and Solar Cells,* vol. 55, pp. 199-214, 1998.

[35] M. Bhuiyan and M. Ali Asgar, "Sizing of a stand-alone photovoltaic power system at Dhaka," *Renewable Energy,* vol. 28, pp. 929-938, 2003.

[36] P. Tsalides and A. Thanailakis, "Loss-of-load probability and related parameters in optimum computer-aided design of stand-alone photovoltaic systems," *Solar Cells,* vol. 18, pp. 115-127, 1986.

[37] B. G. a. A. Ohmura, "The evaluation of surface variations in solar radiation income," *Solar Energy,* vol. 13, pp. 21-34, 1970.

[38] W. Shen, "Optimally sizing of solar array and battery in a standalone photovoltaic system in Malaysia," *Renewable Energy,* vol. 34, pp. 348-352, 2009.

[39] J. Kaldellis, "Optimum technoeconomic energy autonomous photovoltaic solution for remote consumers throughout Greece," *Energy Conversion and Management,* vol. 45, pp. 2745-2760, 2004.

[40] M. Egido and E. Lorenzo, "The sizing of stand alone PV-system: A review and a proposed new method," *Solar Energy Materials and Solar Cells,* vol. 26, pp. 51-69, 1992.

[41] J. Gordon, "Optimal sizing of stand-alone photovoltaic solar power systems," *Solar Cells,* vol. 20, pp. 295-313, 1987.

[42] T. Markvart, A. Fragaki and J. Ross, "PV system sizing using observed time series of solar radiation," *Solar Energy,* vol. 80, pp. 46-50, 2006.

[43] C. Demoulias, "A new simple analytical method for calculating the optimum inverter size in grid-connected PV plants," *Electric Power Systems Research,* vol. 80, pp. 1197-1204, 2010.

[44] P. Arun, R. Banerjee and S. Bandyopadhyay, "Optimum sizing of photovoltaic battery systems incorporating uncertainty through design space approach," *Solar Energy,* vol. 83, pp. 1013-1025, 2009.

[45] A. Petrowski, Evolutionary Algorithms, Wiley, 2017, pp. 3-7.

[46] M. Javadi, A. Jalilvand, R. Noroozian and M. Valizadeh, "Optimal design and economic assessment of battery based stand-alone wind/PV generating system using ABC," in *The 3rd Conference on Thermal Power Plants*, 2011.

[47] S. Singh and S. C. Kaushik, "Optimal sizing of grid integrated hybrid PV-biomass energy system using artificial bee colony algorithm," *IET Renewable Power Generation,* vol. 10, pp. 642-650, 2016.

[48] D. Suchitra, S. Nag and R. Jegatheesan, "Optimal Sizing of Standalone Wind/PV/Fuel Cell/Battery System Using Adaptive Particle Swarm Optimization," *International Journal of Control and Automation,* vol. 9, pp. 327-340, 2016.

[49] A. Mellit, S. Kalogirou, L. Hontoria and S. Shaari, "Artificial intelligence techniques for sizing photovoltaic systems: A review," *Renewable and Sustainable Energy Reviews,* vol. 13, pp. 406-419, 2009.

[50] A. Mellit, S. Kalogirou, L. Hontoria and S. Shaari, "Artificial intelligence techniques for sizing photovoltaic systems: A review," *Renewable and Sustainable Energy Reviews,* vol. 13, pp. 406-419, 2009.

[51] T. Khatib and W. Elmenreich, "An Improved Method for Sizing Standalone Photovoltaic Systems Using Generalized Regression Neural Network," *International Journal of Photoenergy,* vol. 2014, pp. 1-8, 2014.

[52] T. Khatib, A. Mohamed, K. Sopian and M. Mahmoud, "A New Approach for Optimal Sizing of Standalone Photovoltaic Systems," *International Journal of Photoenergy,* vol. 2012, pp. 1-7, 2012.

[53] L. Hontoria, J. Aguilera and P. Zufiria, "A new approach for sizing stand alone photovoltaic systems based in neural networks," *Solar Energy,* vol. 78, pp. 313-319, 2005.

[54] A. Mellit, "ANN-based GA for generating the sizing curve of stand-alone photovoltaic systems," *Advances in Engineering Software,* vol. 41, pp. 687-693, 2010.

Appendices

Appendix A Covariance Matrices of Classification Results

The following six covariance matrix are the PV systems' classification results. Each matrix describes a Gaussian distribution in the features' space. A smaller number in the matrix means a more concentrated distribution. The number in the A.1, A.2, A.3 and A.4 are very small. And these four clusters' application can be observed as mentioned in chapter 3. The values in A.5 is extremely bigger than other covariance matrices. Therefore, the fifth cluster is an outlier's cluster. The visualization of each cluster is shown in Appendix B.

A.1 Covariance Matrix of the First Classification Group

2.02E-02	-2.02E-02	6.93E-05	1.91E-07	4.13E-03	4.95E-05	-1.12E-02	-5.03E-03	-4.09E-03
-2.02E-02	2.04E-02	-1.51E-04	-2.14E-07	-4.16E-03	-9.18E-05	1.13E-02	5.07E-03	4.12E-03
6.93E-05	-1.51E-04	8.12E-05	2.30E-08	3.08E-05	4.22E-05	-1.77E-04	-3.98E-05	-2.63E-05
1.91E-07	-2.14E-07	2.30E-08	5.04E-07	1.07E-07	5.43E-08	-1.85E-07	-7.78E-08	-4.67E-08
4.13E-03	-4.16E-03	3.08E-05	1.07E-07	7.93E-03	-1.31E-05	1.02E-03	-5.64E-04	1.16E-03
4.95E-05	-9.18E-05	4.22E-05	5.43E-08	-1.31E-05	9.04E-05	-1.70E-04	-4.12E-05	-1.57E-05
-1.12E-02	1.13E-02	-1.77E-04	-1.85E-07	1.02E-03	-1.70E-04	2.57E-02	5.95E-03	4.00E-03
-5.03E-03	5.07E-03	-3.98E-05	-7.78E-08	-5.64E-04	-4.12E-05	5.95E-03	3.16E-03	2.13E-03
-4.09E-03	4.12E-03	-2.63E-05	-4.67E-08	1.16E-03	-1.57E-05	4.00E-03	2.13E-03	6.55E-03

Table A.1 Covariance matrix of the first classification group

A.2 Covariance Matrix of the Second Classification Group

1.86E-01	1.66E-04	-1.86E-01	4.77E-02	1.00E-01	-2.10E-01	7.32E-02	3.05E-02	8.40E-02
1.66E-04	9.43E-06	-1.75E-04	1.53E-04	-1.48E-04	-3.76E-04	9.37E-05	-5.49E-05	-3.39E-04
-1.86E-01	-1.75E-04	1.87E-01	-4.79E-02	-1.00E-01	2.10E-01	-7.33E-02	-3.05E-02	-8.36E-02
4.77E-02	1.53E-04	-4.79E-02	1.11E-01	-5.81E-02	9.55E-02	1.56E-02	3.71E-02	8.38E-03
1.00E-01	-1.48E-04	-1.00E-01	-5.81E-02	1.40E-01	-2.33E-01	5.05E-02	1.72E-03	8.54E-02
-2.10E-01	-3.76E-04	2.10E-01	9.55E-02	-2.33E-01	6.07E-01	-5.15E-02	3.02E-02	-8.56E-02
7.32E-02	9.37E-05	-7.33E-02	1.56E-02	5.05E-02	-5.15E-02	8.45E-02	2.70E-02	2.45E-02
3.05E-02	-5.49E-05	-3.05E-02	3.71E-02	1.72E-03	3.02E-02	2.70E-02	3.00E-02	3.90E-02
8.40E-02	-3.39E-04	-8.36E-02	8.38E-03	8.54E-02	-8.56E-02	2.45E-02	3.90E-02	1.45E-01

Table A.2 Covariance matrix of the second classification group

A.3 Covariance Matrix of the Third Classification Group

8.78E-03	8.22E-07	-8.78E-03	7.96E-03	5.30E-05	-3.03E-04	1.06E-04	-5.53E-05	1.24E-04
8.22E-07	1.79E-07	-1.00E-06	2.09E-06	3.36E-07	-7.95E-06	-6.90E-07	-6.07E-07	-3.40E-08
-8.78E-03	-1.00E-06	8.78E-03	-7.96E-03	-5.33E-05	3.11E-04	-1.05E-04	5.60E-05	-1.24E-04
7.96E-03	2.09E-06	-7.96E-03	1.32E-02	-8.90E-04	2.09E-03	-1.11E-03	-1.95E-04	-3.38E-04
5.30E-05	3.36E-07	-5.33E-05	-8.90E-04	1.29E-03	-1.09E-03	8.90E-04	1.77E-04	6.38E-04
-3.03E-04	-7.95E-06	3.11E-04	2.09E-03	-1.09E-03	7.51E-03	-4.18E-04	2.74E-04	-4.15E-04
1.06E-04	-6.90E-07	-1.05E-04	-1.11E-03	8.90E-04	-4.18E-04	1.31E-03	6.09E-04	8.07E-04
-5.53E-05	-6.07E-07	5.60E-05	-1.95E-04	1.77E-04	2.74E-04	6.09E-04	4.98E-04	4.35E-04
1.24E-04	-3.40E-08	-1.24E-04	-3.38E-04	6.38E-04	-4.15E-04	8.07E-04	4.35E-04	6.28E-04

Table A.3 Covariance matrix of the third classification group

A.4 Covariance Matrix of the Fourth Classification Group

5.35E-03	-4.41E-08	-5.35E-03	4.36E-04	2.33E-04	-1.28E-04	-4.33E-07	1.15E-04	2.82E-04
-4.41E-08	6.81E-08	-2.39E-08	-7.01E-07	-3.12E-08	5.30E-07	4.83E-08	5.11E-10	-1.72E-07
-5.35E-03	-2.39E-08	5.35E-03	-4.35E-04	-2.33E-04	1.27E-04	3.85E-07	-1.15E-04	-2.82E-04
4.36E-04	-7.01E-07	-4.35E-04	1.92E-03	-2.61E-04	8.81E-04	-2.89E-05	5.67E-05	6.41E-05
2.33E-04	-3.12E-08	-2.33E-04	-2.61E-04	5.08E-04	1.17E-05	6.36E-05	1.17E-04	2.67E-04
-1.28E-04	5.30E-07	1.27E-04	8.81E-04	1.17E-05	2.97E-03	3.69E-05	3.34E-05	1.62E-05
-4.33E-07	4.83E-08	3.85E-07	-2.89E-05	6.36E-05	3.69E-05	2.25E-05	2.44E-05	3.54E-05
1.15E-04	5.11E-10	-1.15E-04	5.67E-05	1.17E-04	3.34E-05	2.44E-05	1.39E-04	2.38E-04
2.82E-04	-1.72E-07	-2.82E-04	6.41E-05	2.67E-04	1.62E-05	3.54E-05	2.38E-04	5.44E-04

Table A.4 Covariance matrix of the fourth classification group

A.5 Covariance Matrix of the Fifth Classification Group

1.34E+01	-2.52E+00	-1.09E+01	1.33E+01	5.84E+00	5.51E+00	2.98E+01	1.41E+01	1.58E+01
-2.52E+00	2.72E+00	-2.06E-01	-2.95E+00	-4.29E-01	-5.34E+00	-6.02E+00	-1.91E+00	-1.95E+00
-1.09E+01	-2.06E-01	1.11E+01	-1.04E+01	-5.41E+00	-1.66E-01	-2.38E+01	-1.22E+01	-1.39E+01
1.33E+01	-2.95E+00	-1.04E+01	2.07E+01	-2.39E+00	1.80E+01	3.10E+01	1.05E+01	1.27E+01
5.84E+00	-4.29E-01	-5.41E+00	-2.39E+00	2.12E+01	-6.02E+00	1.37E+01	1.49E+01	1.67E+01
5.51E+00	-5.34E+00	-1.66E-01	1.80E+01	-6.02E+00	3.35E+01	1.60E+01	2.14E+00	4.05E+00
2.98E+01	-6.02E+00	-2.38E+01	3.10E+01	1.37E+01	1.60E+01	6.86E+01	3.32E+01	3.74E+01
1.41E+01	-1.91E+00	-1.22E+01	1.05E+01	1.49E+01	2.14E+00	3.32E+01	4.03E+01	3.71E+01
1.58E+01	-1.95E+00	-1.39E+01	1.27E+01	1.67E+01	4.05E+00	3.74E+01	3.71E+01	3.61E+01

Table A.5 Covariance matrix of the fifth classification group

A.6 Covariance Matrix of the Sixth Classification Group

1.46E-01	-7.65E-02	-6.95E-02	-2.39E-03	8.03E-02	-7.89E-02	7.16E-02	-4.52E-03	4.29E-02
-7.65E-02	5.68E-02	1.97E-02	5.66E-04	-2.58E-02	2.09E-02	-2.61E-02	1.25E-02	-3.74E-03
-6.95E-02	1.97E-02	4.98E-02	1.83E-03	-5.45E-02	5.80E-02	-4.55E-02	-7.99E-03	-3.91E-02
-2.39E-03	5.66E-04	1.83E-03	7.70E-05	-1.63E-03	2.19E-03	-1.34E-03	-2.07E-04	-1.06E-03
8.03E-02	-2.58E-02	-5.45E-02	-1.63E-03	1.90E-01	-6.39E-02	1.95E-01	5.01E-02	1.13E-01
-7.89E-02	2.09E-02	5.80E-02	2.19E-03	-6.39E-02	6.90E-02	-5.44E-02	-1.03E-02	-4.53E-02
7.16E-02	-2.61E-02	-4.55E-02	-1.34E-03	1.95E-01	-5.44E-02	2.20E-01	6.45E-02	9.26E-02
-4.52E-03	1.25E-02	-7.99E-03	-2.07E-04	5.01E-02	-1.03E-02	6.45E-02	4.08E-02	2.32E-02
4.29E-02	-3.74E-03	-3.91E-02	-1.06E-03	1.13E-01	-4.53E-02	9.26E-02	2.32E-02	1.19E-01

Table A.6 Covariance matrix of the sixth classification group

Appendix B Scatter Plots of Classification Result

Each cluster is plotted in different scatter plot matrix. By observing the features' distribution and their correlation, we can infer their operation strategy and then determine the appropriate sizing method. The visualization process can also verify the classification results and see if the results distinguish systems in the database clearly.

B.1 Scatter Plots of the First Classification Group

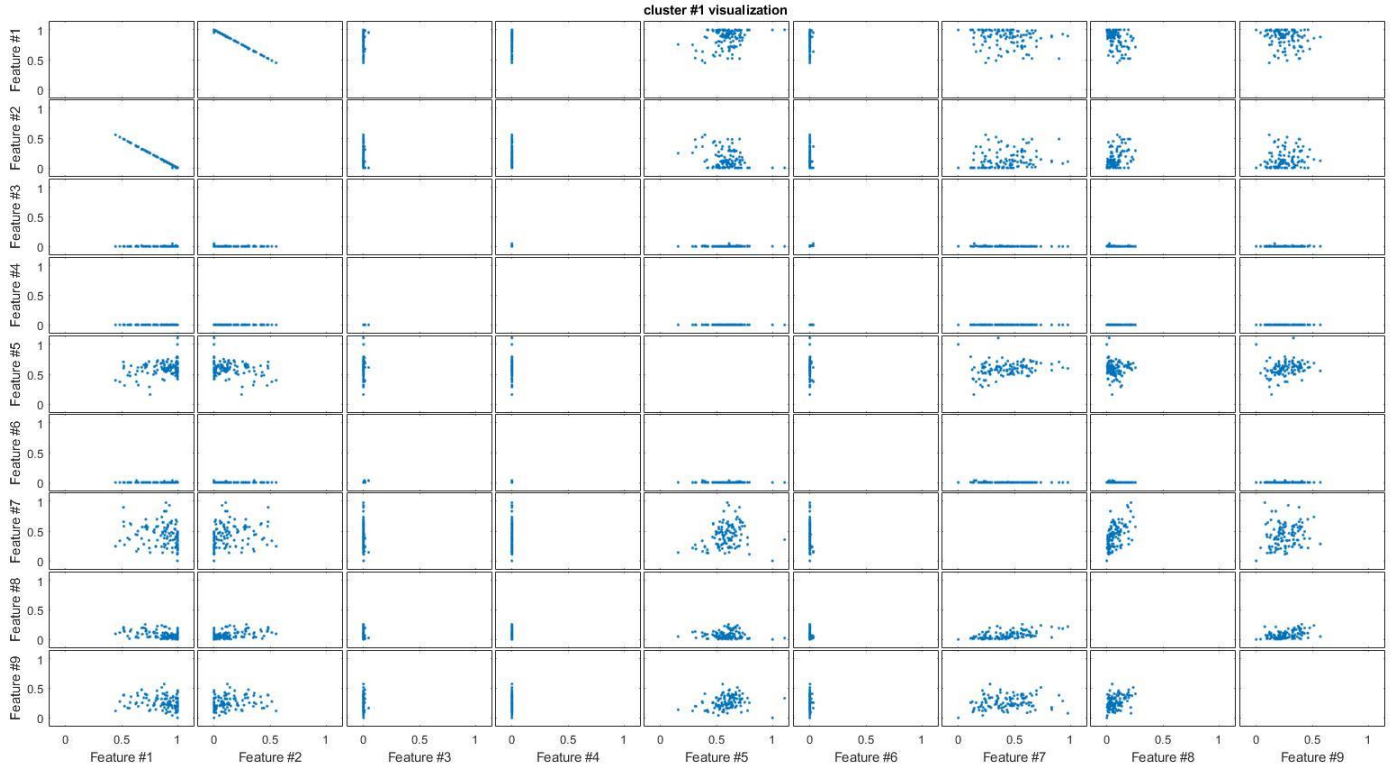

Table B.1 Scatter plots of the first classification group

In the type-one systems, PV and diesel generator provides all required energy. They do not have any interaction with the grid, so they are stand-alone PV/diesel hybrid systems. This type of system tends to use more PV than the diesel generator, probably due to low operational cost. Moreover, about 50% of the energy charges their battery bank and used when PV is not available. So, we can expect more battery degradation in this type of system. We find some system uses a diesel generator for a large portion of energy generation. They may have cheap diesel sources but more likely for other reasons. When the PV panel is undersized, the generated energy is not enough for the load demand and diesel generator will start and provide the deficit energy. When the battery is undersized, the stored energy cannot satisfy load demand at night, and the diesel generator will be activated. To determine the reason, we need to know the SoC data, but this type of data is not available. Another observation is that feature #7 and #9 are widely distributed, and they do not correlate with other features. It may be due to the variety of their local climates.

B.2 Scatter Plots of the Second Classification Group

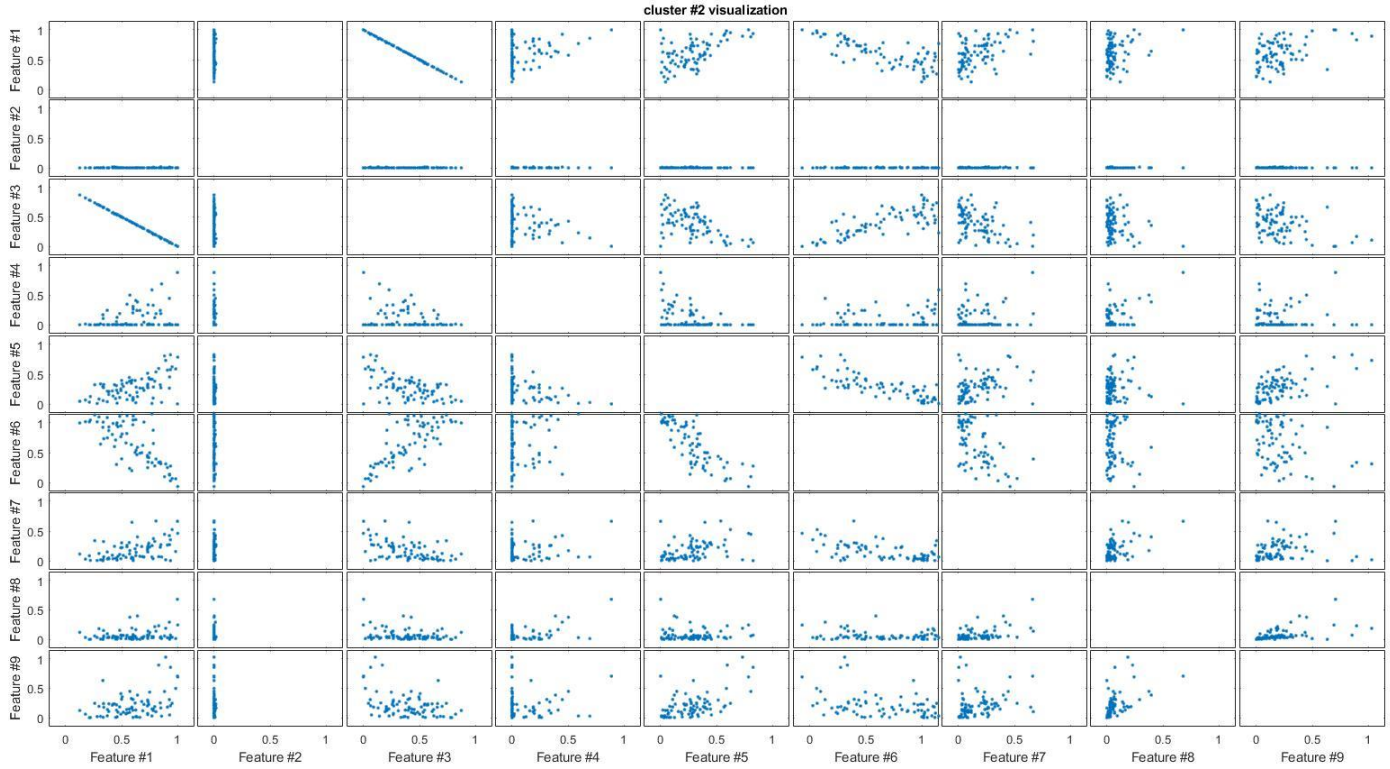

Table B.2 Scatter plots of the second classification group

The type-two systems are grid-tied because the systems purchase energy from the grid (feature #3) or sell to the grid (feature #4). Feature #1 is distributed widely in the range of [0,1], which indicate that some system heavily relies on the PV energy, but some are not. We notice that feature #1, #3, #5 and #6 are correlated. The system which takes energy from the grid at night will also use less energy from the battery and lightly relies on PV energy. This observation may be due to the constraints of the available area for PV panels. We must consider such constraints in the optimization. When we observe the feature#8, we notice some systems have higher values. These systems may use battery energy during the peak hours to reduce the grid power and shave the peak.

B.3 Scatter Plots of the Third Classification Group

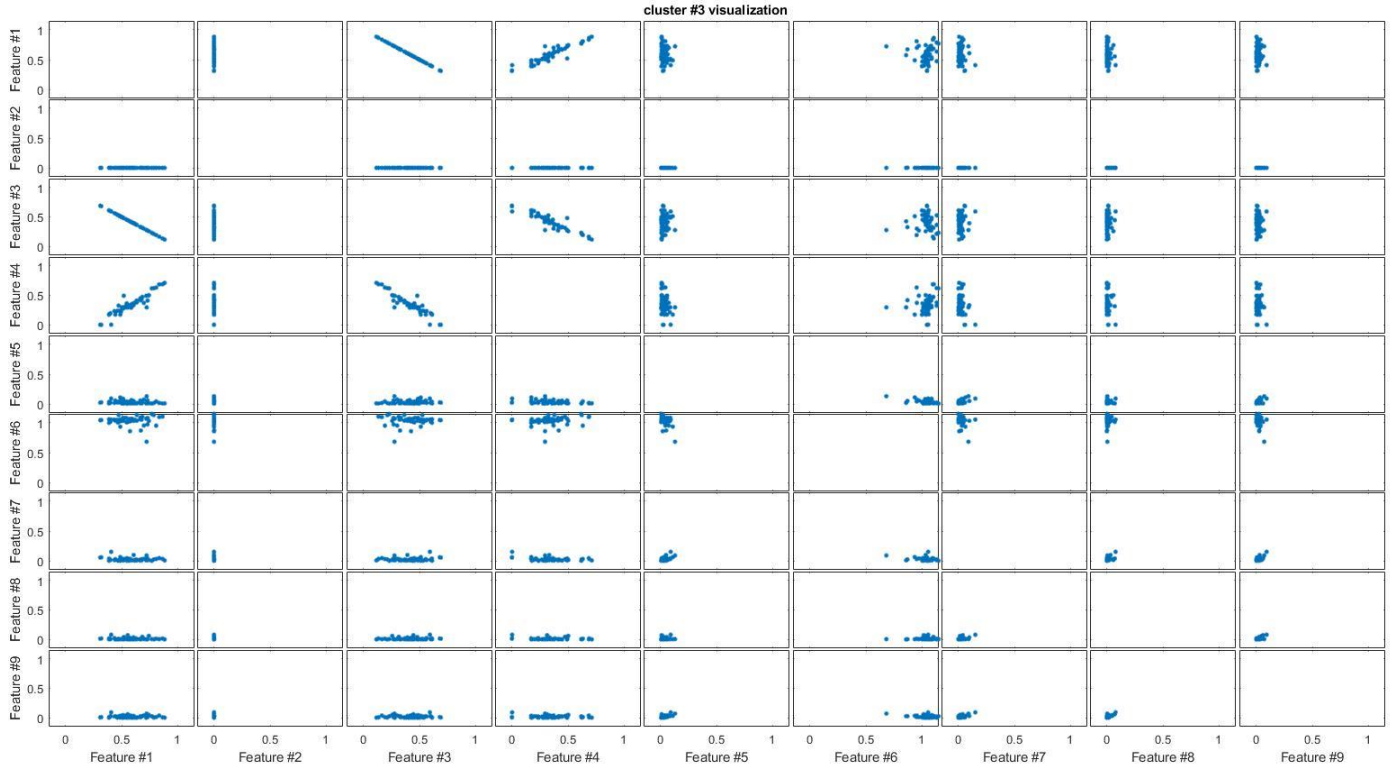

Table B.3 Scatter plots of the third classification group

B.4 Scatter Plots of the Fourth Classification Group

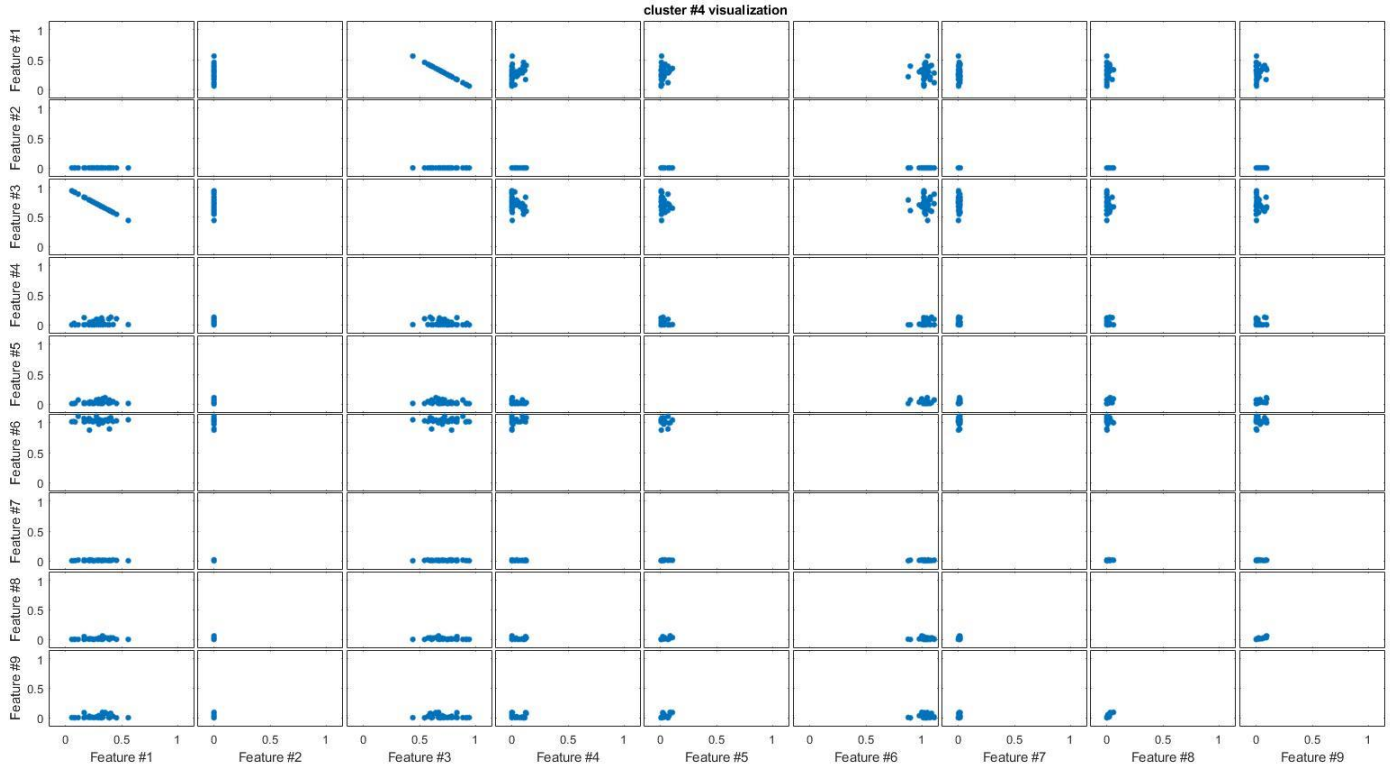

Table B.4 Scatter plots of the fourth classification group

The third and fourth groups' features are concentrated, which implies they are well classified. The two types of systems are the uninterruptable power supply because of their batteries neither charge nor discharge in a day. Type-three and type-four systems are different in feature #4. The type-three systems inject surplus PV energy into the connected grid. A UPS, like a backup source, drives their loads during the grid outage. To size these systems, we must consider the grid reliability and how long the UPS can withstand.

B.5 Scatter Plots of the Fifth Classification Group

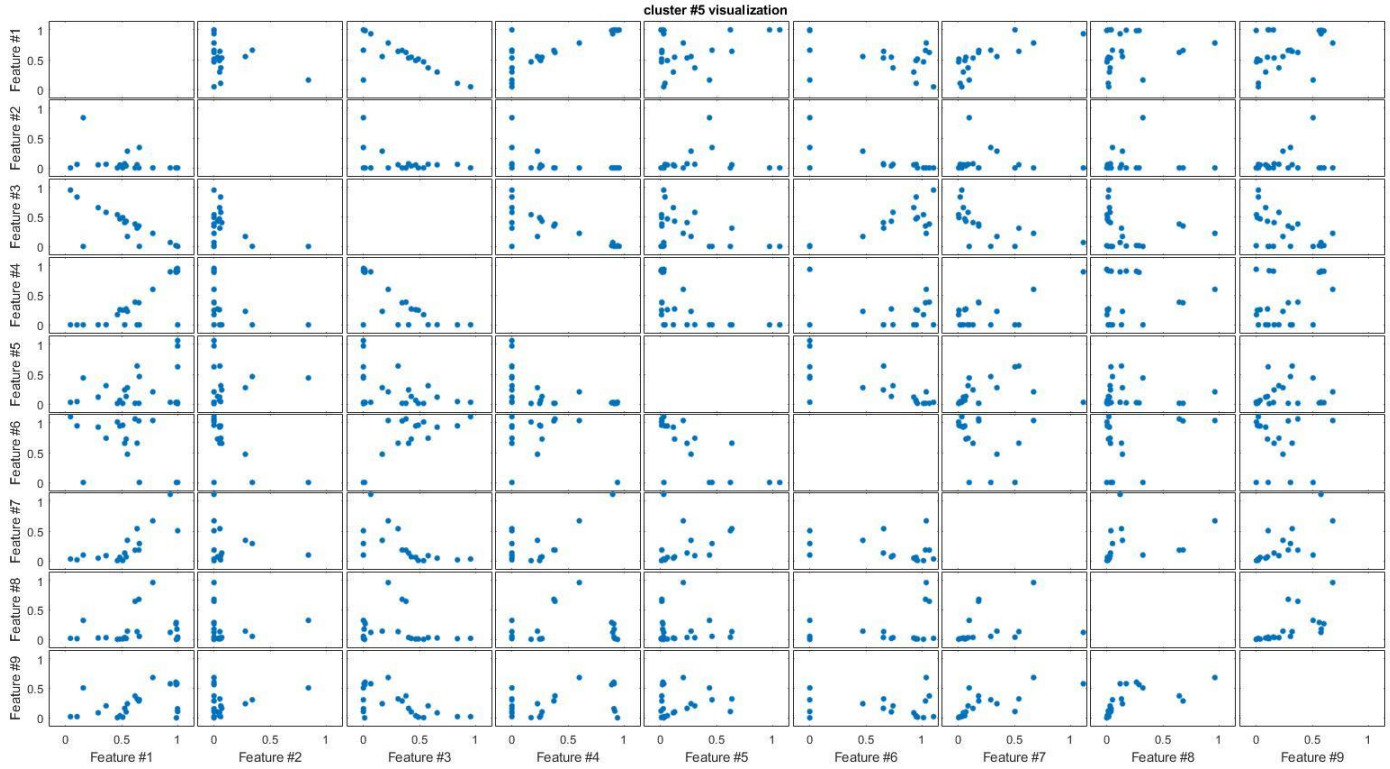

Table B.5 Scatter plots of the fifth classification group

The fifth class seems a group for outliers. All nine features should be in the range of [0,1] by their definition. In the covariance matrix of the fifth GMM component, most variances and covariances are bigger than one. It indicates that the component covers most spaces. It more likely contains outliers.

B.6 Scatter Plots of the Sixth Classification Group

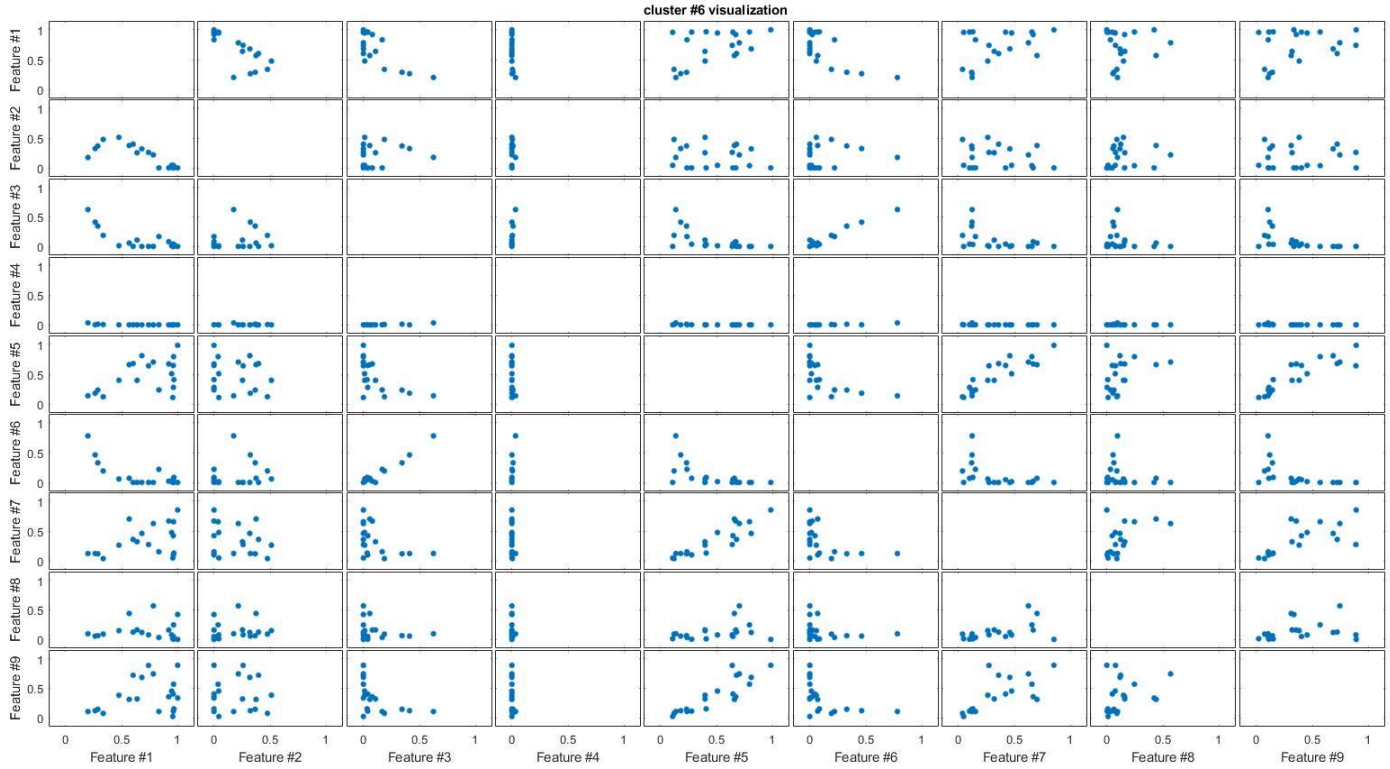

Table B.6 Scatter plots of the sixth classification group

The sixth group's members are spread widely in the feature space. It seems a group for outliers like the fifth class. However, the model's covariance and variance value are small. It still indicates some specialties over other groups. My observation is that this group contains all diesel/PV grid-tied system. However, in the database, the number of such systems are very small. So, the classification result wrongly classifies a few systems into this group.

Appendix C Artificial Bee Colony Method

In the section 1.4.2, it is shown that the numerical method is widely used for optimizing battery size and PV panel size for a stand-alone PV system based on load and illumination profiles. However, in the given database, there are many types of PV system and each system may have more than two variables to be optimized. The complexity of the numerical method grows exponentially with respect to the increment of variables. Therefore, the ABC is proposed to solve the problem.

For example, to optimize a PV/ battery/ diesel generator hybrid stand-alone system, the numerical method needs to calculate the LLP with respect to the combination of PV array size (S_{pv}), battery size (S_{batt}) and the power rating of the diesel generator (S_{gen}). The calculation of LLP relies on a simulation based on the historical data. LLP($S_{pv}, S_{batt}, S_{gen}$) represents the calculation result. There are a great number of possible combinations such as (1, 1, 1), (1, 1, 2) … (1, 1, $S_{max\ gen}$) … (1, 2, $S_{max\ gen}$), (1, 3, $S_{max\ gen}$), …, (1, $S_{max\ batt}$, $S_{max\ gen}$), …, ($S_{max\ pv}$, $S_{max\ batt}$, $S_{max\ gen}$). And the huge amount makes the work very complexed.

On the contrary, the ABC method requires an objective function. In this case, the objective function is the overall cost of the system.

$$\text{Cost}(S_{pv}, S_{batt}, S_{gen})$$

The ABC method random the initial value of $S_{pv}, S_{batt}, S_{gen}$ and makes perturbation for variable after each iteration until the optimization result achieved.

The ABC algorithm is mimicking bee colony to optimize mathematical function. It can deal with non-leaner functions and constraints. A colony consists of three types of bees:

employed bees, onlooker bees, and scout bees. The number of employed bees, $\boldsymbol{SN}$, needs to be set. Bee's locations, $\boldsymbol{X_i}$ ($\boldsymbol{i = 1, 2, 3 \ldots SN}$), are possible solutions of the optimization problem. Each location is a $\boldsymbol{D}$-dimensional vector. Value of the optimization problem can be calculated based on the given locations. The calculated values are the amount of the nectar, which also called fitness (fit_i). Each bee updates a new location and compare the nectar at the new location with the old one. If the new location has a better food source, the bee will memorize the new location, forget the old location, and share their best location with the whole colony. Each bee follows different rule to update their new location according to the bee's type. The employed bees search the location around their old location. The onlooker bees tend to move to the locations has more nectar according to the shared information in the colony. The scout bees search randomly within the searching space. The employed bees and onlooker bees utilize the similar formula to generate the new location,

$$\boldsymbol{V_{i,j} = X_{i,j} + \Phi_{i,j}\left(X_{i,j} - X_{k,j}\right)} \qquad (1)$$

where $k \in \{1, 2, \ldots, \boldsymbol{SN}\}$ and $j \in \{1, 2, \ldots, \boldsymbol{D}\}$.$\boldsymbol{V_{i,j}}$ is one coordinate of new location. $\boldsymbol{X_{i,j}}$ is the old location's coordinate. $\boldsymbol{\Phi_{i,j}}$ is a random number between [-1, 1]. The employed bees chose *k* randomly, but the onlooker bees chose *k* location which has more nectar. The probability of choosing k location can be expressed as,

$$p_k = \frac{fit_k}{\sum_{n=1}^{SN} fit_n} \qquad (2)$$

In the algorithm, a limit value need to be set as the limit for abandonment. If a bee's location does not change after the predetermined number of cycles. The bee will behave as a scout bee. It will move to a random location without capering with the old food source. Assume

the abandoned location is $\boldsymbol{X_i}$ and $j \in \{1, 2, \ldots, \boldsymbol{D}\}$. The operation of randomizing next location can be defined as:

$$\boldsymbol{X_i^j} = \boldsymbol{X_{min}^j} + \text{rand}(0, 1)(\boldsymbol{X_{max}^j} - \boldsymbol{X_{min}^j}) \qquad (3)$$

To run the algorithm, there values should be set: The number of employed bees ($\boldsymbol{SN}$), the limit cycles for food source abandonment and the maximum cycling number ($\boldsymbol{MCN}$). The detailed ABC algorithm is given below:

1. Initialize the population of solutions, $\boldsymbol{X_{i,j}}$, $\boldsymbol{i = 1, 2, 3 \ldots SN}$, $\boldsymbol{j = 1, 2, 3 \ldots D}$
2. Evaluate the fit_i for the whole population
3. cycle=1
4. **repeat**
5. Produce new solutions $\boldsymbol{V_{i,j}}$for the employed bees by using (1) and evaluate the fitness of the new solutions
6. Compare the fitness of the old solution and the new solution. Keep the solutions with higher fitness
7. Calculate the probability values p_i of the solutions $\boldsymbol{X_i}$ by (2)
8. Produce the new solutions $\boldsymbol{V_{i,j}}$ for the onlookers from the solutions $\boldsymbol{X_{i,j}}$ selected depending on p_i and evaluate their fitness
9. Compare the fitness of the old solution and the new solution. Keep the solutions with higher fitness
10. Check if any bee stays at the same solution longer than the limit cycles for food source abandonment. if exists, abandon it and replace it with a new randomly produced solution $\boldsymbol{X_{i,j}}$ by (3)
11. Memorize the best solution of the whole colony
12. cycle=cycle+1
13. **until** cycle= $\boldsymbol{MCN}$

Three types of bee follow different rules updating their next solution. For the employed bees and onlooker bees, when they generate the new solutions, a random number between [-1, 1] participate the evaluation and give the algorithm certain level of randomness. But the variation is proportional to the distance between two bees, $\boldsymbol{X_{i,j} - X_{k,j}}$. When the two types of bee close to each other, the difference of $\boldsymbol{X_{i,j}, X_{k,j}}$ could be very small. Without the help of scout bees, the

whole colony may be trapped at a local optimum solution. Therefore, the scout bee and limit cycles for food source abandonment are necessary.

As discussed above, the ABC algorithm keeps generating new locations for employed bees and comparing the new locations with the old locations. The algorithm always memorizes the locations with better fitness. It is called greedy rule. In order to solve constrained problems, the ABC algorithm replaces the greedy rule with Deb's method. The Deb's method consists three steps: 1) Any solution within constraints is preferred to an infeasible solution, 2) Among two feasible solutions, the one with better fitness is preferred, 3) Among two infeasible solutions, the one violating constrains least is preferred.

In the studied database, year-round power flows information is available. The climate at a specific location is unlikely to change dramatically. And the usage behavior of a system is also assumed to be similar year by year. In this case, the last-year power follows information can be summarized as a year-round profile and used for system optimization.

To optimize sizes for battery, PV panel and a diesel generator in a system, an objective function need to be formed based on the system's purpose and limitation. The objective function can be a cost function in most cases and the function can be very complicated to includes different types of cost of all components in the system. The function can be too complicated to use the mathematical method. The numerical method is able to solve non-linear functions and tolerant to the change of the objective functions.

The optimization problem brings challenges and the optimization method need to be chosen based on the challenges. We may optimize multiple variables together. The method is better to find the global optimal solution rather than trapped by a locally optimal solution. It must deal with constraints caused by physical limitation. The ABC algorithm satisfied all requirements and, therefore, chosen to optimize such a problem.

Appendix D Sizing Results of the Second Case Study

The following table shows the results of fifty sizing process for the second case study described in Section 5.5. Because randomness is the sizing method, the fifty trials give different sizing results. However, their results and annual costs are similar.

	Size of PV panel (kW)	Size of battery (kWh)	Size of diesel generator (kW)	Annual cost ($)
Trial 1	2.959907	21.42524	0.045997	1459.568
Trial 2	2.984417	21.36711	0.046133	1459.242
Trial 3	2.983289	21.36899	0.04658	1458.998
Trial 4	2.984696	21.37994	0.046756	1459.696
Trial 5	2.984507	21.36814	0.04615	1459.288
Trial 6	2.983854	21.36732	0.047233	1459.367
Trial 7	2.983474	21.37277	0.046151	1459.291
Trial 8	2.960632	21.42309	0.046742	1459.45
Trial 9	2.988146	21.35838	0.046521	1459.415
Trial 10	2.958392	21.42942	0.04618	1459.359
Trial 11	2.983484	21.37251	0.04726	1459.533
Trial 12	2.987209	21.36372	0.046788	1459.503
Trial 13	2.983546	21.37468	0.046665	1459.273
Trial 14	2.98791	21.36049	0.047036	1459.627
Trial 15	2.986266	21.37015	0.046212	1459.634
Trial 16	2.985265	21.36875	0.046615	1459.329
Trial 17	2.985312	21.36722	0.046747	1459.303
Trial 18	2.984123	21.38088	0.046655	1459.612
Trial 19	2.984049	21.38247	0.046726	1459.675
Trial 20	2.983467	21.37111	0.046262	1459.17
Trial 21	2.983789	21.37437	0.04644	1459.301
Trial 22	2.986201	21.36235	0.046283	1459.287
Trial 23	2.986118	21.36835	0.046893	1459.541
Trial 24	2.961824	21.42397	0.046457	1459.654
Trial 25	2.987711	21.35937	0.046804	1459.425
Trial 26	2.984644	21.36728	0.046427	1459.173
Trial 27	2.985293	21.36528	0.046555	1459.195
Trial 28	2.986296	21.36389	0.046307	1459.355
Trial 29	2.983645	21.36782	0.046649	1459.02

Trial 30	2.957862	21.43115	0.047159	1459.552
Trial 31	2.984109	21.36871	0.046473	1459.131
Trial 32	2.98409	21.37058	0.046351	1459.225
Trial 33	2.984989	21.36451	0.047116	1459.345
Trial 34	2.957931	21.42941	0.046846	1459.276
Trial 35	2.983353	21.37048	0.04641	1459.081
Trial 36	2.985285	21.36947	0.046484	1459.36
Trial 37	2.983694	21.36892	0.046779	1459.102
Trial 38	2.961738	21.42052	0.046505	1459.502
Trial 39	2.982809	21.37338	0.046953	1459.204
Trial 40	2.98733	21.35957	0.046493	1459.323
Trial 41	2.985094	21.36741	0.046812	1459.294
Trial 42	2.9856	21.36513	0.046591	1459.243
Trial 43	2.98398	21.37035	0.04655	1459.169
Trial 44	2.984061	21.36918	0.046104	1459.28
Trial 45	2.983928	21.37251	0.046502	1459.245
Trial 46	2.958114	21.44224	0.046351	1459.749
Trial 47	2.987776	21.36055	0.04635	1459.464
Trial 48	2.983056	21.37034	0.046735	1459.035
Trial 49	2.986171	21.36539	0.047005	1459.5
Trial 50	2.958373	21.43686	0.0464	1459.574

Appendix E Gaussian Mixture Model

Previous studies have proposed various classification method for the multivariate time series data. And the method is significantly relying on nature of the given data. It is necessary to consider how to maximize the similarity between the within-group objects, meanwhile, minimize the similarity between the between-group objects. Therefore, based on the possible applications mentioned in section 1.3, seven features are proposed to distinguish PV systems. Ideally, the data's dimension is reduced from a few hundred to seven while keeping their differences. In the proposed procedures, GMM method is implemented during the classification process because GMM method will provide the classification in round shape, allow overlaps, and have the potentiality to classify outliers into a group.

The Gaussian Mixture Model (GMM) is a distribution-based clustering method. A GMM is the sum of M components describing the distribution density of N samples. A model can fit any distribution by increasing the number of components. Each component is described as a Gaussian distribution function with corresponding weight, mean value and covariance matrix. A multivariate Gaussian probability density function is defined as

$$(\mathrm{x}|\theta) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum|^{\frac{1}{2}}}\exp(-\frac{(x-\mu)^T\sum^{\frac{1}{2}}(x-\mu)}{2})$$

Where $\theta = (\mu, \sum)$, D is the model's dimension, μ is the mean vector and $\sum$ is the covariance matrix. Then, the GMM can be defined as

$$\mathrm{P}(\mathrm{x}|\theta) = \sum_{k=1}^{M} \alpha_k \varphi(\mathrm{x}|\theta_k)$$

Where k is the number of components in the GMM, α is the weight of the component and $\varphi(x|\theta_k)$ the probability density function. By definition, $\alpha > 0$ and $\sum_{k=1}^{M} \alpha_k = 1$.

Regarding a GMM, we can apply maximum likelihood method to estimate the weigh, covariance matrix and mean value for each component in the model. Its log-likelihood is given:

$$\log L(\alpha, \theta_k) = \sum_{j=1}^{N} \log(\sum_{k=1}^{M} \alpha_k \varphi(x_j|\theta_k))$$

Then the function is a sum of logarithm functions, contains hidden variables (the cluster each sample belonging to), and contains many unknown parameters $\alpha, \mu, \sum$. Therefore, the function must be solved through iterations.

Expectation-Maximization algorithm is an iterative method to solve the likelihood maximization problem who contains hidden variables. At the beginning of the algorithm, parameters $\alpha, \mu, \sum$ are initialized randomly. In each iteration, it contains two steps: the expectation step and the maximization step.

1) E-step

According to the value of $\alpha, \mu, \sum$, the process calculates the probability of sample j belonging to model component k.

$$\gamma_{jk} = \frac{\alpha_k \varphi(x_j|\theta_k)}{\sum_{k=1}^{M} \alpha_k \varphi(x_j|\theta_k)}$$

where $k \in \{1, 2, \dots, \boldsymbol{M}\}$ and $j \in \{1, 2, \dots, \boldsymbol{N}\}$.

2) M-step

Once all expectations are calculated, the GMM parameters can be updated.

$$\mu_k = \frac{\sum_{j=1}^{N}(\gamma_{jk} x_j)}{\sum_{j=1}^{N} \gamma_{jk}}$$

$$\sum_k = \frac{\sum_{j=1}^{N} \gamma_{jk}(x_j - \mu_k)(x_j - \mu_k)^T}{\sum_{j=1}^{N} \gamma_{jk}}$$

$$\alpha_k = \frac{\sum_{j=1}^{N} \gamma_{jk}}{N}$$

The algorithm repeats E-steps and M steps until all parameters converge. Then the GMM of the dataset is formed. For any new sample x, we can repeat the E-step to calculate the probabilities. The new sample will be assigned to the cluster with higher probability.

The database has been recording power flows of more than 4000 sites. Each site may have different devices and run for a different purpose. Therefore, to optimize the device sizes of a system, it is important to understand the operation purpose of the site. For example, a grid-tied system may use a battery bank as a back-up source. Usually, only a small amount of energy cycles in the battery every day. But the site needs a big battery to prevent grid outage. In another case. If a site is stand-alone and has its own diesel generator, the battery only participates in the energy cycling. To optimize the battery size in such a system, all economic effects should be considered.