POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Elettrica

A Method to Classify Substation Load Profiles Based on PCA

Relatore: Prof. Alberto Berizzi

Correlatore: Dott. Alessandro Bosisio

Tesi di Laurea Magistrale di:
Tianxing Qi
Matr. 10534003

Anno Accademico 2017-2018

INDEX

INDEX OF FIGURE

INDEX OF TABLE

ABSTRACT

Abstract in italiano

Questa tesi tratta un metodo per classificare i profili di carico delle cabine secondarie MT, basato sull'analisi delle componenti principali (PCA).

La procedura di classificazione si articola su cinque passaggi: 1) Caricamento dei profili di carico: in questa fase la base dati viene omogeneizzata e semplificata. 2) Preparazione dei dati: questa fase include quattro processi: identificazione e ed eliminazione delle serie non valide, normalizzazione dei dati, interpolazione dei dati e rilevamento dei valori anomali. Una combinazione di Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) e Autoregressive Moving-Average (ARMA) viene applicato per l'interpolazione dei dati agli estremi della serie. 3) Semplificazione dei dati tramite PCA. La PCA che viene utilizzata per estrarre le componenti principali, al fine di ridurre la dimensione dati base. 4) Classificazione dei profili di carico: le cabine secondarie vengono classificate in 8 gruppi sulla base delle prime 3 componenti principali 5) Interpretazione dei risultati.

Analizzando i risultati della classificazione, si può ipotizzare che: 1) Le prime 3 componenti principali rappresentano: il fattore presenza/assenza di attivita, il fattore di condizionamento estivo e il fattore carico feriale/festivo. 2) Definendo 8 tipologie di profilo, solo quattro risultano significative.

I clienti di "tipo 1" sono caratterizzati da una uniforme attivita durante tutto l'anno, forte utilizzo del condizionamento estivo, limitata differenza tra carico feriale e festivo; i clienti di "tipo 2" sono caratterizzati da una riduzione dell'attivita durante il periodo agostano, forte utilizzo del condizionamento estivo, limitata differenza tra carico feriale e festivo; i clienti di "tipo 3" catatterizzati da una riduzione dell'attivita durante il periodo agostano, basso utilizo del condizionamento estivo, sensibile differenza tra carico feriale e festivo; i clienti di "tipo 4" caratterizzati da una uniforme attivita durante tutto l'anno, basso utilizzo del condizionamento estivo, sensibile differenza tra carico feriale e festive.

Abstract in inglese

This thesis is about a method to classify MV/LV substation load profiles based on principle component analysis (PCA) method, considering the real daily load data from the UNARETI company in Milan. The classification model includes five steps: 1) Loading raw load profiles: in this step time series are refined and simplified. 2) Data preparation: this step includes four processes: defining and deleting the bad data, normalizing data, the data interpolation and outlier detecting. A method that combines PCHIP and ARMA is applied to deal with the boundary issues. 3) Data simplifying with PCA method: PCA method is used to extract the main components, in order to decrease the dimension of the data. 4) Load profile classification: MV/LV substations are classified in 8 groups based on the first 3 components. 5) Result applications. By analyzing the load customers classifying in Milan, it can be concluded that 1) The main 3 components of daily load profile are: the activity factor, the air conditioner factor and the holiday factor. 2) After the classification, 8 types of substations have been identified but only four result to be common. Type 1 customers have the character of uniform activities, high airconditioner occupancy, low holiday differences, for example commercial center; Type 2 customers have the character of non-uniform activities, high airconditioner occupancy, low holiday differences; Type 3 customers have the character of non-uniform activities, low airconditioner occupancy, high holiday differences; Type 4 customers have the character of uniform activities, low airconditioner occupancy, high holiday differences.

Keyword: load classification; principle components analysis; data processing; load profile

1 INTRODUCTION

Load classification has always been the basis of power system planning, peak management, flexible electricity price, and load forecasting. Good load classification methods can provide correct basis and guidance for system planning and peak load management. Therefore, a method of load classification is studied in this thesis. Based on real data, this method is applied to the load classification of the Milano distribution, and the load characteristics of the power grid are analyzed in detail.

1.1 Research Background

Nowadays, the load is growing rapidly, but the analysis of power load characteristics is still at a relatively less studied stage, mainly reflected in the following aspects:

(1) The models and data required for the analysis of load characteristics are dispersed in multiple production and management systems. The underlying data has not been integrated and unified data management and analysis cannot be performed;

(2) There are many types of users in the power system, and various types of users exhibit different load characteristics. Currently, there is a lack of a scientific and effective load classification method and a comprehensive load characteristic index system that meets the actual conditions of the power grid.

(3) There are many kinds of influencing factors for the classification of load characteristics, and the influences for each kind are different. At present, there is no in-depth and systematic analysis for the factors affecting the change of load characteristics and the degree of influence.

(4) The research on the classification load characteristics is not thorough, and its rules are not accurately studied. An effective statistics and analysis system cannot be formed. The effective technical support and guidance can not be provided for load forecasting, power grid planning, economic dispatch and electricity market.

The above problems have restricted the further improvement of the load management and application in the power grid, and lead to the difficulties to adapt to the requirements of refined management and technical progress of the power grid. At present, there are

many researches on power grid load forecasting, but there are relatively few studies on load classification. Load classification is the basis of load forecasting. Through the analysis of load characteristics and load classification, it is very important to understand the changes and trends of power grid load. It is necessary, therefore, to study accurate and appropriate load classification method is of great significance.

In the long run, accurate load classification methods are not only beneficial to the power system, but also beneficial to users, and are specifically expressed in the following aspects:

(1) Beneficial to power system

The scientific and accurate load classification method can save the country's infrastructure investment in the power industry, improve the thermal efficiency of power generation equipment, reduce fuel consumption, reduce the cost of power generation, increase the safety and stability of power system operation, improve the quality of power supply, and is conducive to the overhaul of power equipment. At the same time, it is an important basis for power planning, production, and operation. It is also an important reference for formulating relevant policies. It provides technical guidelines for the production and operation in power grids, power grid planning, the power grids accurate management, and innovative creation work. To meet the requirements of the development and improvement of the economy, power companies must understand the market, rely on the market, open up markets, and scientifically accurate load classification to optimize the planning and operation of the power grid, improve the quality of customer service, and ensure the safety, economy, quality and environmental protection of the power grid.

(2) Beneficial to the users

The ultimate beneficiaries should be mostly users. Since the country's investment in user equipment can be saved, shaving peak load and filling valley load, where electricity consumption during peak hours is led to the valley hours, can reduce electricity bills, which in turn reduces production costs and benefit the people in urban and rural residents. As a result of the measures taken, employees of factories and factories have taken a break from work and staggered work shift times, so that the load of service industries such as local transportation, water supply and gas supply and so on can be balanced.

1.2 Research status

With the development of smart grid, more and more smart meters are installed into distribution networks [1]. Consumption behaviors of customers are known through load curves data collected from smart meters. Clustering technology is very useful for data mining in smart gird. In competitive electricity market with severe uncertainties, performing effective customer classification according to customers' electrical behavior is important for setting up new tariff offers [2].

The load classification is to separate enormous load profiles into several typical clusters.[3] In recent years, researchers have proposed a variety of clustering methods. Various methods for clustering load curves have been used in the load clustering in recent years. such as K-means[4]-[5], fuzzy c-means(FCM) [6]-[8], hierarchical methods[9], [10], self-organizing map (SOM) [11], support vector machine (SVM) [12], [13], subspace projection Method [14].

In the meanwhile, there are some papers combining some methods together. In [15], a method combine with hierarchical and fuzzy-c mean was introduced. [16] proposes a two-stage clustering algorithm combining supervised learning methods to classify electric customer.

With the development of data mining technologies, some new clustering methods have emerged for electricity consumption patterns classification. In paper [17], they built a prediction model to identify the customers who would most likely respond to the prospective offering of the company. Thus in [18], the extreme learning machine method is used to analyze the nontechnical loss to classify the customers.

However, in the data mining technologies, instead of improving the method of clustering, decreasing the dimension of the data and extract the main features of the load profile can be another important method to cluster. In [19], a statistical analysis of end-users’ historical consumption is conducted to better capture their consumption regularity. With the features captured, it is easier to classify the customers. [20] focus on the description of the construction and implementation of the recognition of customers' risk preference model. Consider the following information: customer consignment streamline data, customer exchange streamline data, customer fund streamline data, and so on. Using

PCA, two important indicators, behavioral characteristics and risk preferences of the customers, are selected, with which the construction of the customer classification model is built with K-means clustering algorithm.

1.3 Thesis content and outline

According to the achievements and existing problems in the field of power load classification, this thesis analyzes daily load data of more than 4000 MV/LV substations in Milano. A methodology to classify distribution load profiles based on PCA method is put forward. Based on the basic theory of classification of load characteristics, a procedure has been defined as follow:

(1) Select the substation daily peak load curve (366 values, in the year 2016) as the load feature vector of the power system user, and deal with the data in three steps: defining bad data, data interpolation and dealing with outliers;

(2) PCA (Principal Component Analysis) is used to analyze the load eigenvectors of the substations, the principal components are extracted, the dimension of the data set is reduced, and each principal component is analyzed.

(3) According to the results of the principal component analysis, MV/LV substations load profile are classified, and combined with geographical information, a set of load types are defined and classified.

The thesis is organized as follows:

(1) The basic concepts and theories of load classification. This chapter mainly introduces three aspects: the load characteristics and type of power system, the extraction and preprocessing of load data, and the basic load classification method.

(2) Basic concepts of load profile classifying. This chapter mainly introduces the basic concept of load classification. On one side, it introduced the profile classification already known and the need to produce a new load classification. On another side, the load data preprocessing, which is one of the most important task in load classification, is studied in four aspects, that is bad data processing, load curve normalization, missing value processing, and outlier processing.

(3) The basic concepts and theories of principal component analysis. This chapter introduced the basic principles and properties of principal component analysis, and lists the basic steps of principal component analysis. In addition, two important points of principal component analysis are discussed in the end of the chapter: data requirements and the data standardization.

(4) The load classifying model based on PCA method. This chapter mainly introduced the five steps of the procedure, that is: loading raw load data from the database, load data preparation, data simplifying with PCA method, load profile classification and applications. Based on the real data from the Milano distribution network, among all the five steps, the loading of raw data is mainly introduced.

(5) The result analysis. In this chapter, at first, the result of PCA is given. And after the analysis, the main components are settled down. Then, with the result of PCA, the classification of the substations is given, together with the meaning of each classification.

2 BASIC CONCEPTS OF LOAD PROFILE CLASSIFYING

With the in-depth development of the electricity market and the extensive application of DSM technology, the load classification of power systems has become a very important basic work, such as load pricing, load forecasting, system planning, and load modeling. To complete all of the work needs to have a scientific and accurate classification of the type of load. Therefore, the in-depth study of methods and applications of power system load classification helps to timely master the changes in the law and trends of power load, and it helps to the scientific management of electricity load, which is beneficial to planning for the development of electricity use. Therefore, it has important theoretical and practical significance.

2.1 Load classification in power system

The basic task of the power system is to provide uninterrupted high-quality electrical power to meet the needs of various types of loads. Normally, the load is the workload that a transformer substation or a grid is responsible for at a certain moment. For the user, the power load refers to the sum of the power consumed by all the user's devices connected to the power grid at a certain moment.

The power system load characteristics, whether for the system planning and design, or the optimized safety operation, are extremely important. Therefore, research on load characteristics is an important task of the power system. Research on load forecasting, load classification, load pricing, load demand side management and load modeling has become an important research area in the modern power market environment, and it is also an important content in the field of power system automation research.

The actual power system has a variety of loads, and the total system load is a sum of various types of loads. Therefore, for the actual power system, all types of loads should have a formal and accurate classification, so that the same type of load has the same or similar characteristics, thereby simplifying or reducing the difficulty and complexity of load management.

In general, the load can be divided into urban civilian loads, commercial loads, rural loads, industrial loads, and other loads. The urban civilian load mainly refers to the urban household load, and the commercial load and industrial load are the load of commercial and industrial services. The rural load refers to the load of all rural areas (including the rural civilian load, production, irrigation and commercial electricity, etc.), while other loads include municipal electricity (e.g. street lighting), utilities, government offices, railways and trams, military and others. According to different standards, different types of loads can be divided:

1. Divided by physical performance

According to the physical performance of the load can be divided into active load and reactive load.

(1) Active load: It is the energy that is converted into other forms of energy and is actually consumed in electrical equipment. The unit of calculation is kW (kilowatts).

(2) Reactive power load: In the process of power transmission and conversion, it is necessary to establish the power consumed by magnetic fields (such as transformers, motors, etc.) and electric fields (such as capacitive field energy). It only completes the conversion of electromagnetic energy and does not do work. In this sense, it is called "reactive power" and the unit of calculation is l kvar.

2. Divided by electrical energy

According to the production, supply, and production of electricity, the load can be divided into generation load, power supply load, and electricity load.

(1) Power generation load: It refers to the power load that the power plant supplys to the power grid, plus the power load consumed by the power plant itself at the same time, and the unit is kW.

(2) Power supply load: It refers to the sum of the power generation load of each power plant in the power supply area, minus the power load consumed by the power plant itself, plus the load inputted from other power supply area, minus the load absorbed by the other power supply area, unit kW.

(3) Customer load: It refers to power supply load in the area minus the loss in the line and the transformer. The unit of calculation is kW.

3. Divided by time scale

The load can be divided into yearly load, monthly load, daily load and hourly load .

4. Divided by requirements for reliability of power supply

According to the nature of the electricity load and the different requirements for the reliability of the power supply, the load can be divided into level-one load, level-two load and level-three load.

(1) Level-one load: The event of a power outage in this load will result in personal injury or death, or political, military, and economically significant losses, such as an accident that endangers personal safety, causes unrecoverable damage or disorder to key equipment in industrial production, resulting in major losses in the national economy. The loads of important military facilities, hospitals, airports and other places generally belong to this class of loads.

(2) Level-two load: When the blackout occurs in this load, it will result in production cuts, work stoppages, traffic congestion in some areas, and troubles in the normal activities for a large number of residents in cities. For instance, the load of residents and factories in large and medium-sized cities generally are always classified into level-two load.

(3) Level-three load: It is general loads other than level-one and level-two loads. The loss caused by the interruption of such loads is not significant. Such as the factory's ancillary workshops, small towns and rural public loads, etc. are often classified as such loads.

5. The international general classification of electricity load

According to the international general classification of the power load, the power load can be divided as follows:

(1) Agriculture, forestry, animal husbandry, fishery and water conservancy. Rural irrigation, agricultural side production, agriculture, forestry, animal husbandry, fisheries, water conservancy and other kinds of load are included.

(2) Industry. A variety of mine industries, manufacturing and other kinds of load are included.

(3) Geological survey and exploration industry.

(4) Construction industry.

(5) Transportation, post and telecommunications. Loads in road and railway station, ship terminals, airport, pipeline transportation, electrified railway and post and telecommunication are included.

(6) Commercial, public catering, material supply and warehousing, various stores, catering, material supply units, and warehouse loads.

(7) Other institutions. The load in the city's public transport, street lighting, art and sports institutions, national party and government agencies, various social groups, welfare services, scientific research institutions and other units is included.

(8) Urban and rural household electricity.

Although load in the power system can be divided according to the above criteria, such classification is not strict and accurate. In the actual classification, it may happen that the actual load is reduced to what kind of dispute.

In this case, it is generally determined by the power supply department itself. Therefore, in some power supply departments, each of them may have its own more load classification list in detail in order to prepare for load classification. However, there are still some problems in the practical application of the load classification method adopted by the above power supply department:

(1) Users in the same work field may have different load characteristics.

At present, the power supply department classifies the power system user load, most of which is based on the industry to which the user load belongs and the economic activity characteristics of the user. However, as the composition of the user's load devices becomes more and more complex and people's production and lifestyle changes, the user's load characteristics in the same industry are not exactly the same, and their load curves may have large differences.

(2) It cannot reflect changes and differences in the power grid.

With the development of economy and society, there will be some new types of user loads in the power grid, and these types of user loads may have large differences from the already defined types of user loads. Therefore, it is necessary to reconsider the type and definition. And there are also certain differences in the load composition between the power grids in different regions. The load types of the power grids should be divided according to the actual situation.

(3) Inaccurate classification affects the further application on this basis.

The traditional load classification method does not fully consider the actual characteristics and rules of the customer load, lacks theoretical basis. And it reduces the accuracy and rationality of the classification result, affecting some applications. For example, it reduces the accuracy of the classification load forecasting, which causes the unreasonable electricity price policy and so on.

Therefore, in order to solve the above problems, it is necessary to study a scientific and accurate load classification method to provide powerful reference and basis for load classification based on which new applications can be applied in the power supply department.

2.2 Load data preprocessing

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data. Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected. Despite the amount of time it takes, there has been surprisingly little research on how to clean data well. Part of the challenge is the breadth of activities it encompasses: from outlier checking, to date parsing, to missing value imputation.

The load database is extremely large, and the database can be easily affected by noise, lost data, and inconsistent data. Low quality load data will lead to low accuracy of classification results, so data preprocessing is an important first step in load classification. Data preprocessing usually includes four aspects: bad data processing, load curve normalization, missing value processing, and outlier processing.

In power systems, almost all research on loads is based on raw data. Therefore, the accuracy of the research results is determined by the correctness of the original data. The raw data is usually directly derived from the real-time data collected in the EMS/SCADA system. Due to the dynamic data acquisition, there are sometimes channel failures, congestion, and other phenomena. In addition, the interruption of the data acquisition program can also cause errors in the original data. At the same time, according to the needs of the classification method used, the data needs to be normalized and other processing. Therefore, before applying the method of this thesis to study the system load classification, we need to preprocess the sample data from the following aspects.

2.2.1 Bad data processing

Bad data can be defined in just a few general classifications. The vast majority of data quality professionals would break data quality issues in these terms:

1. Incomplete data: A data record that is missing the necessary values that render the record at least partially useless.

2. Invalid data: A data record that is complete, but has values that are not within pre-established parameters.

3. Inconsistent data: Data that is not entered in the same way as the other data records.

4. Unique data: A data record unlike any other previously measured data record.

5. Legacy Data: Defining data as too old can also be very important in order to utilize data that is still relevant to the organization, and excluding data that was collected before certain variables were established.

There are many causes of bad data, which may be determined by the inherent characteristics of the data generation mechanism, or may be due to imperfect data acquisition equipment, data transmission errors, data loss and other human controllable factors. Bad data includes missing values, 0 values, and straight line load values. Bad data identification includes physical identification and statistical identification.

The physical identification method is based on people's experience in identifying bad data. The statistical identification method is to give a confidence probability and determine

a confidence threshold. Any error that exceeds this threshold is considered not to belong to the random error range, and it is considered as a method of rejecting bad data. The physical identification is usually used in bad load data detection.

2.2.2 Normalization

The user load data obtained through the power system load measurement device will have a large difference in the value range, and these differences will have a great impact on the classification result. Therefore, sample data should be normalized before classification to eliminate the influence of these differences.

There are several ways to normalize data. Here introduces five normalization methods:

To make it simple, the original user load is presented as $X = (x_1, x_2, ..., x_n)$.

1. Feature scaling

The user data is normalized according to Equation (2-1), and the value x_i is mapped to x_i' in the interval [a,b]. The feature scaling maintains the relationship between the original data values.

$$x_i' = \frac{x_i - \min(X)}{\max(X) - \min(X)}(b-a) + a \qquad i = 1, 2, ..., n \tag{2-1}$$

$$\max(X) = \max(x_1, x_2, ..., x_n) \tag{2-2}$$

$$\min(X) = \min(x_1, x_2, ..., x_n) \tag{2-3}$$

Normally, a = 0 and b = 1, so the normalized data range is between 0-1.

2. Feature scaling standardization

After the feature scaling standardization, each user's load has a mean value of 0, with a range difference of 1, and $\left|x_i'\right| < 1$. So the error can be reduced in the PCA.

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{2-4}$$

$$x_i' = \frac{x_i - \overline{X}}{\max(X) - \min(X)} \qquad i = 1, 2, ..., n \tag{2-5}$$

3. Student's t-statistic standardization

After the student's t-statistic standardization, the average load of each user is 0 and the standard deviation is 1. The equation shown below is the calculation:

$$x_i' = \frac{x_i - \overline{X}}{s}, i = 1, 2, ..., n \tag{2-6}$$

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \tag{2-7}$$

It is always used to normalize residuals when population parameters are unknown

4. Fractional scaling normalization

The decimal scaling is normalized by moving the decimal places in x_i. The number of decimal places to move depends on the maximum absolute value of x_i. The formula is as follows, where i is the smallest integer which makes $\max(|x_i'|) < 1$.

$$x_i' = \frac{x_i}{10^i} \tag{2-8}$$

In this method, all the load can be restricted to [0,1]. Due to the factor is the base of 10, it is easier to reconstruct the original data intuitively. And the shape of data is saved. But it has a problem that the maximum load data of each case is not fixed.

5. Max load normalization

The max load normalization is normalized by dividing the maximum load. This normalization is based on the fact that all the load data are larger than zero. This kind of normalization is simple, and the load shape can be reserved.

$$x_i' = \frac{x_i}{\max(X)}, i = 1, 2, ..., n \tag{2-9}$$

The table 2.1 shows the different properties of these normalization methods.

Table 2.1 Typical normalization methods

Feature scaling	Maintains the relationship between the original data values
Feature scaling standardization	Each user's load has a mean value of 0, with a range difference of 1
Student's t-statistic standardization	The average load of each user is 0 and the standard deviation is 1
Fractional scaling normalization	Easier to reconstruct the original data intuitively; Shape is saved; the maximum load data of each case is not fixed
Max load normalization	The maximum load data of each case is not fixed; Shape is saved

2.2.3 Missing value processing

During the data collection process, due to equipment or machinery, some missing values can be generated. Dealing with missing values of non-bad data, interpolation is an effective method. In the field of numerical analysis, interpolation is a method of constructing new data points within the range of a discrete set of known data points.

A closely related problem is the approximation of a complicated function by a simple function. Suppose the formula for some given function is known, but too complicated to evaluate efficiently. A few data points from the original function can be interpolated to produce a simpler function which is still fairly close to the original. The resulting gain in simplicity may outweigh the loss from interpolation error.

Given the values of an unknown function f(x) in table 2.2, there are some methods that can be used to interpolate the missing data：

Table 2.2 Value table of unknown function f(x)

x	f(x)
0	0
1	0.8415
2	0.9093
3	0.1411
4	-0.7568
5	-0.9589
6	-0.2794

1. Piecewise constant interpolation

The simplest interpolation method is to locate the nearest data value, and assign the same value. In simple problems, this method is unlikely to be used, as linear interpolation is almost as easy, but in higher-dimensional multivariate interpolation, this could be a favorable choice for its speed and simplicity.

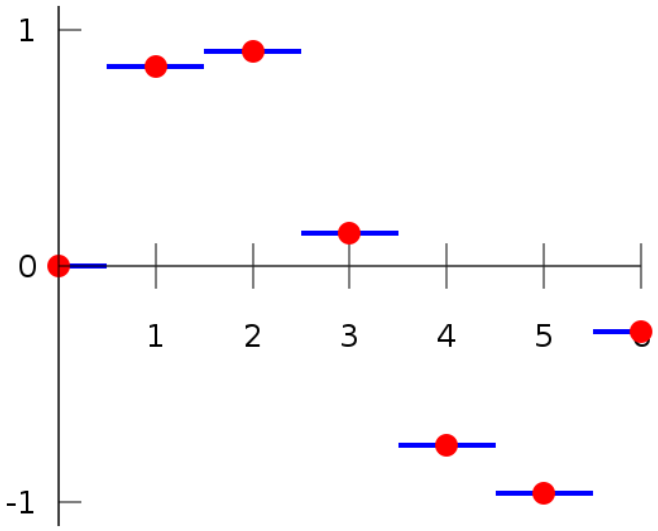


Figure 2.1: Example of the piecewise constant interpolation

2. Linear interpolation

Another simple method is linear interpolation (sometimes known as lerp). Consider the above example of estimating f(2.5). Since 2.5 is midway between 2 and 3, it is reasonable to take f(2.5) midway between f(2) = 0.9093 and f(3) = 0.1411, which yields 0.5252.

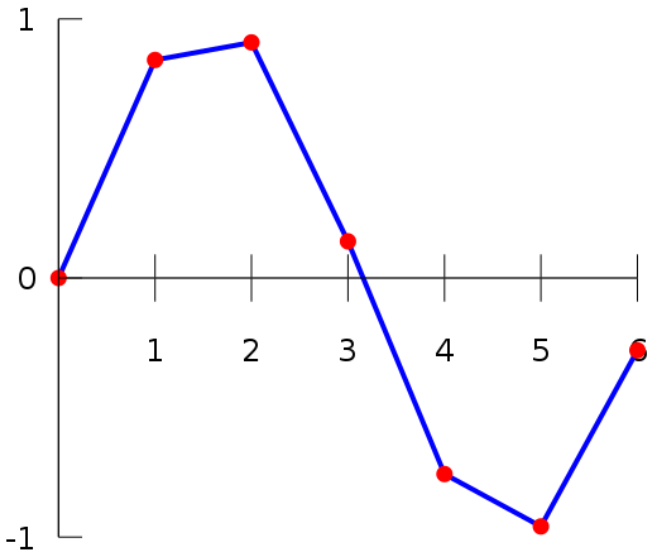


Figure 2.2: Example of the linear interpolation

Generally, linear interpolation takes two data points, say (x_a, y_a) and (x_b, y_b), and the interpolation at the point (x, y) is given by:

$$y = y_a + (y_b - y_a)\frac{x - x_a}{x_b - x_a} \tag{2-10}$$

This previous equation states that the slope of the new line between (x_a, y_a) and (x, y) is the same as the slope of the line between (x_a, y_a) and (x_b, y_b).

Linear interpolation is quick and easy, but it is not very precise. Another disadvantage is that the interpolation is not differentiable at the point x_k.

The following error estimate shows that linear interpolation is not very precise. Denote the function which we want to interpolate by g, and suppose that x lies between x_a and x_b and that g is twice continuously differentiable, then the linear interpolation error is:

$$\left| f(x) - g(x) \right| \le C(x_b - x_a)^2 \tag{2-11}$$

Where

$$C = \frac{1}{8} \max_{r \in [x_a, x_b]} \left| g''(r) \right| \tag{2-12}$$

Thus the error is proportional to the square of the distance between the data points. The error in some other methods, including polynomial interpolation and spline interpolation, is proportional to higher powers of the distance between the data points. These methods also produce smoother interpolation.

3. Spline interpolation

Spline interpolation uses low-degree polynomials in each of the intervals, and chooses the polynomial pieces such that they fit smoothly together. The resulting function is called a spline.

For instance, the natural cubic spline is piecewise cubic and twice continuously differentiable. Furthermore, its second derivative is zero at the end points. The natural cubic spline interpolating the points in the table above is given by

$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1] \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2] \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3] \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4] \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5] \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6] \end{cases} \tag{2-13}$$

In this example we get f(2.5) = 0.5972.

Like polynomial interpolation, spline interpolation incurs a smaller error than linear interpolation and the interpolation is smoother. However, the interpolation is easier to evaluate than the high-degree polynomials used in polynomial interpolation. Even the global nature of the basis functions leads to ill-conditioning.

4. Piecewise Cubic Hermite Data Interpolation(PCHIP)

In numerical analysis, a cubic Hermite spline or cubic Hermite interpolator is a spline where each piece is a third-degree polynomial specified in Hermite form: that is, by its values and first derivatives at the end points of the corresponding domain interval.

Cubic Hermite splines are typically used for interpolation of numeric data specified at given argument values $x_1,x_2,\ldots,x_n$, to obtain a smooth continuous function. The data should consist of the desired function value and derivative at each x_k. If only the values are provided, the derivatives must be estimated from them. The Hermite formula is applied to each interval (x_k, x_{k+1}) separately. The resulting spline will be continuous and will have continuous first derivative.

Given a set of interpolation points $x_0<x_1<\ldots<x_N$, the Piecewise Cubic Hermite Spline Interpolation(PCHIP), S to the function f, satisfies:

(i) $S \in C^1[x_0, x_N]$

(ii) $S(x_i) = f(x_i)$ and $S'(x_i) = f'(x_i)$ for i=0,1,…,N.

(iii) On each interval $[x_k, x_{k+1}]$, S is a cubic polynomial.

We can construct these splines as follows. On each interval $[x_k, x_{k+1}]$, let S be the cubic polynomial S_i is given by:

$$S_i(x) = c_0 + c_1(x - x_{i-1}) + c_2(x - x_{i-1})^2 + c_3(x - x_{i-1})^3 \qquad (2\text{-}14)$$

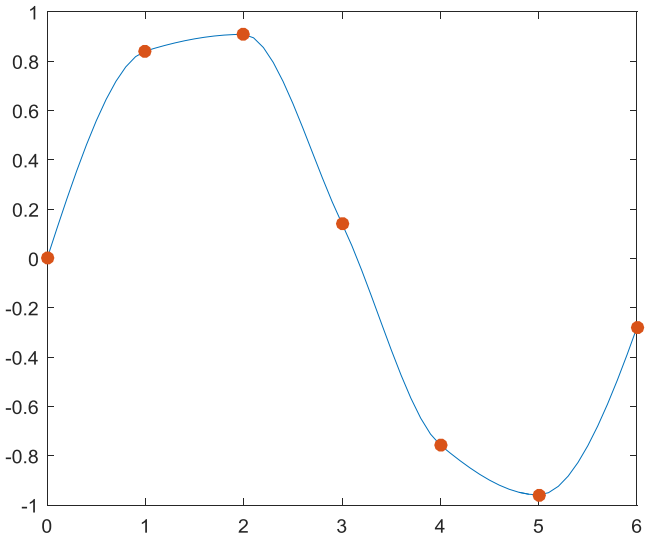


Figure 2.3: Example of Piecewise Cubic Hermite Data Interpolation

The figure 2.3 shows the interpolation in the previous case. PCHIP method features software to produce a monotone and "visually pleasing" interpolation to monotone data. Such an interpolation may be more reasonable than a cubic spline if the data contain both 'steep' and 'flat' sections.

5. PCHIP plus ARMA method

PCHIP is an excellent method to keep the shape of the curves when the data contain both 'steep' and 'flat' sections. But there is a deadly weakness: it cannot deal with the boundary interpolation. When there is some amount of lacking data on the boundary, it is not accurate to use PCHIP method to interpolate. In the meanwhile, it will bring the deadly disorder due to the property of the cubic interpolation.

For instance, if we want to interpolate until 10 in the example, the interpolation shows out as figure 2.4:

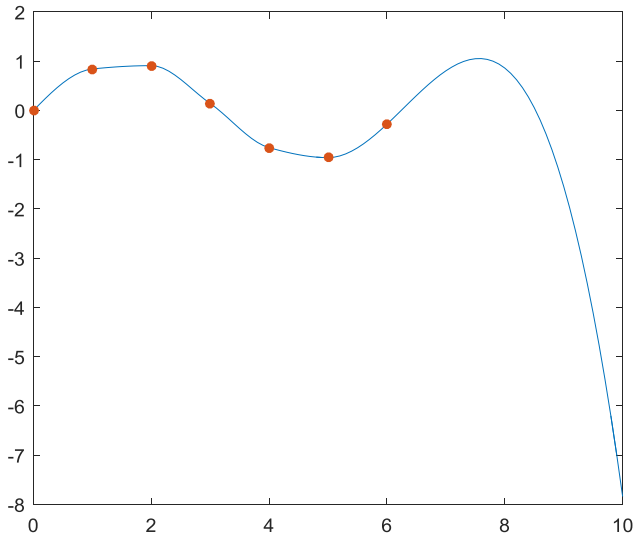


Figure 2.4: The example of boundary problem with PCHIP method

To overcome this issue, we used a PCHIP plus ARMA(autoregressive–moving-average) method.

In the statistical analysis of time series, ARMA models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the autoregression and the second for the moving average.

Given a time series of data X_t , the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The model consists of two parts: the autoregressive (AR) part and the moving average (MA) part. The (AR) part involves regressing the variable on its own lagged (i.e., past) values. The (MA) part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past.

The model is usually referred to as the ARMA(p,q) model where p is the order of the autoregressive part and q is the order of the moving average part.

1. Autoregressive model

The notation AR(p) refers to the autoregressive model of order p. The AR(p) model is written

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{2-15}$$

where $\varphi_1, \ldots, \varphi_p$ are parameters, c is a constant, and the random variable ε_t is white noise.

Some constraints are necessary on the values of the parameters so that the model remains stationary. For example, processes in the AR(1) model with $|\phi_1| \geq 1$ are not stationary.

2. Moving-average model

The notation MA(q) refers to the moving average model of order q:

$$X_t = \mu + \sum_{i=1}^{q} \theta_i X_{t-i} + \varepsilon_t \tag{2-16}$$

where the $\theta_1, \ldots, \theta_q$ are the parameters of the model, μ is the expectation of X_t (often assumed to equal 0), and the $\varepsilon_t, \varepsilon_{t-1}, \ldots$ are again, white noise error terms.

3. ARMA model

The notation ARMA(p, q) refers to the model with p autoregressive terms and q moving-average terms. This model contains the AR(p) and MA(q) models,

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i X_{t-i} + \sum_{i=1}^{p} \varphi_i X_{t-i} \tag{2-17}$$

The general ARMA model was described in the 1951 by Peter Whittle, who used mathematical analysis (Laurent series and Fourier analysis) and statistical inference. ARMA models were popularized by a 1970 book by George E. P. Box and Jenkins[22], who expounded an iterative (Box–Jenkins) method for choosing and estimating them. This method was useful for low-order polynomials (of degree three or less).

The figure 2.5 shows the PCHIP plus ARMA method:

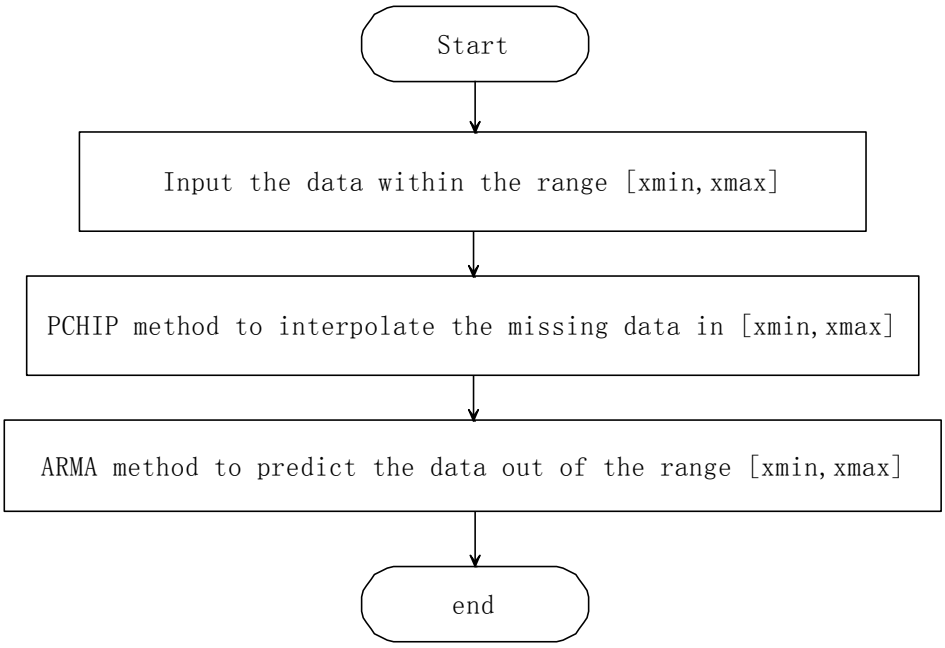
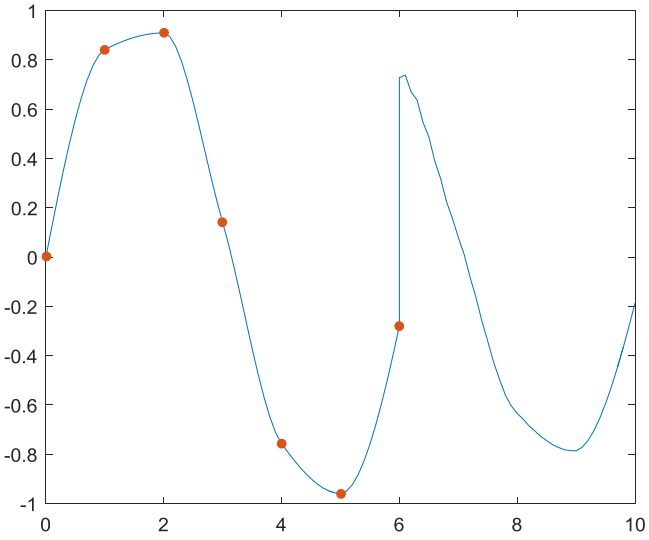


Figure 2.5: The PCHIP plus ARMA interpolation method process

The figure 2.6 shows the result of the interpolation using PCHIP plus ARMA methd. Compared with the result of PCHIP method, the PCHIP plus ARMA methodis able to better capture features of the original data set.



\

Figure 2.6: The example of PCHIP plus ARMA interpolation in boader

2.2.4 Outliers processing

In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental

error. The latter sometimes have to be excluded from the data set. An outlier can cause serious problems in statistical analysis.

In larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected.

There is no rigid mathematical definition of what constitutes an outlier, determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection. Some are graphical such as normal probability plots, others are model-based or hybrid.

Considering that normally distributed data are common in the reality, here the outlier detection based on 3-σ rule is introduced.

In the case of normally distributed data, the 3-σ rule means that roughly 1 in 22 observations will differ by twice the standard deviation or more from the mean, and 1 in 370 will deviate by three times the standard deviation.[21] In a sample of 1000 observations, the presence of up to five observations deviating from the mean by more than three times the standard deviation is within the range of what can be expected, being less than twice the expected number and hence within 1 standard deviation of the expected number and not indicate an anomaly. If the sample size is only 100, however, just three such outliers are already reason for concern, being more than 11 times the expected number. The figure 2.7 shows the probability of the normally distributed data.

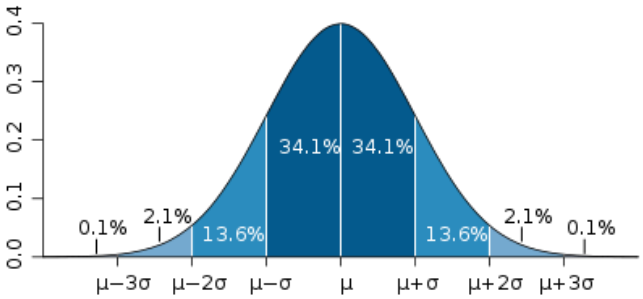


Figure 2.7: The possibility distribution for normal distributed data

Thus, as the equation shown below, 3- σ rule is used in the outlier detection in daily load. The daily load can be signed as $X = (x_1, x_2, ..., x_n)$ with the mean value $\overline{X}$ and the standard deviation σ, then we say that x_i is an outlier if $x_i > \overline{X} + 3\sigma$ or $x_i < \overline{X} - 3\sigma$.

2.3 Summary

In this chapter, the basic concept of load profile classifying and data processing is introduced. On one hand, the basic definition of load is introduced, together with five general load classifications: the load classification divided by physical performance, electrical energy, time scale, reliability and the international classification nowadays. On the other hand, the concept of load data preprocessing is introduced. According to the problems existing in the data preparing, the methods in the bad data processing, the data normalization, the missing value processing and the outliers processing are introduced. Especially in the interpolation of missing values, a method that combine PCHIP and ARMA interpolation is put forward to deal with the boundary problem.

3 BASIC CONCEPT OF PCA

3.1 Introduction of PCA

Principal component analysis (PCA) is an important statistical method to study how to convert multiple-index problems into fewer comprehensive indicators. Those few comprehensive indicators are not related to each other and can provide most of the information of the original indicators. PCA transforms high-dimensional space problems into lower dimensional one.

In addition to reducing the dimensions of multivariate data systems, PCA also simplifies the statistical characteristics of variable systems. PCA can provide many important system information, such as the location of the gravity center (or average level) of the data points, the maximum direction of data variation and the distribution range of the group points.

As one of the most important multivariate statistical methods, PCA has its applications in various fields such as social economy, enterprise management, and geology, biochemistry engineering. In comprehensive evaluation, process control and diagnosis, data compression, signal processing, model recognition and other directions, PCA method has been widely used.

3.2 Basic idea

The basic idea of PCA can be summarized as follows:

With an orthogonal transformation, the component-dependent original random variable is converted into a new uncorrelated variable. From an algebraic perspective, the covariance matrix of the original variable is converted into a diagonal matrix. From the geometric point of view, the original variable system is transformed into a new orthogonal system, so that it points to the orthogonal direction in which the sample points are dispersed, and then the dimensionality of the multidimensional variable system is reduced. In terms of feature extraction, PCA is equivalent to an extraction method based on the minimum mean square error.

3.2.1 Goal

The following considerations leads people to introduce PCA method when dealing with big data:[23]

(1) The number of components should be significantly less than the original number of data, preferably only one;

(2) The newly constructed components should reflect as much as possible the behavior and various information of the project under the original data;

(3) If there is more than one component, they should not be related to each other, and different components have different degrees of importance

3.2.2 Definition and property

Definition 1: Assume $X=(X_1,X_2,...,X_p)$ is a p-dimensional random vector whose *i* th principal component can represent as $Y_i=u_i'X, i=1,2,...,p$, Where u_i is *i* th column vector of the orthogonal matrix U, and satisfies the following conditions:

(1) Y_i is the linear combination of $X_1,X_2,...,X_p$ with the largest variance;

(2) Y_k is irrelevant to $Y_1,Y_2,...,Y_{k-1}$, where *k*=2,3,…,*p*.

Property 1: Assume that $\sum$ is the covariance matrix of the random vector $X=(X_1,X_2,...,X_p)$, and its eigenvalues-eigenvectors pairs can be $(\lambda_1,e_1),(\lambda_2,e_2),...,(\lambda_p,e_p)$, where $\lambda_1\geq\lambda_2\geq...\geq\lambda_p\geq 0$. So the *i* th component can be:

$$Y_i=e_i'X=e_{i1}X_1+e_{i2}X_2+...+e_{ip}X_p, i=1,2,...,p \tag{3-1}$$

Where

$$\mathrm{var}(Y_i)=e_i'\sum e_i=\lambda_i, i=1,2,...,p \tag{3-2}$$

$$\mathrm{cov}(Y_i,Y_k)=e_i'\sum e_k=0, i\neq k \tag{3-3}$$

Property 2: Assume that random vector $X=(X_1,X_2,...,X_p)$ has a covariance matrix $\sum$, and its eigenvalues-eigenvectors pairs can be $(\lambda_1,e_1),(\lambda_2,e_2),...,(\lambda_p,e_p)$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Y_k is the main component, then:

$$\sigma_{11}+\sigma_{22}+...+\sigma_{pp}=\sum_{i=1}^{p}\mathrm{var}(X_i)=\lambda_1+\lambda_2+...+\lambda_p=\sum_{i=1}^{p}\mathrm{var}(Y_i) \qquad (3\text{-}4)$$

The property 2 shows that the covariance matrix $\sum$ of the principal component vectors is the diagonal matrix Λ. The variance represents the variability and reflects the amount of information. It is easy to know from the above analysis that the total variance= $\sigma_{11}+\sigma_{22}+...+\sigma_{pp}=\lambda_1+\lambda_2+...+\lambda_p$, that is the total amount of information can be measured by eigenvalues. The corresponding eigenvalues reflect the amount of information corresponding to the principal components. [24]

This leads to the following definition:

Definition 2: $\dfrac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ is the contribution rate for the k principal component, $\dfrac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$ is the cumulative variance contribution rate of the first k principal components.

Property 3: If $Y_i=e_i'X$ is the principal component obtained from the covariance matrix $\sum$, then:

$$\rho(Y_k,X_i)=\frac{e_{ki}\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} \quad i,k=1,2,...,p \qquad (3\text{-}5)$$

is the correlation coefficient between component Y_k and variable X_i, where $(\lambda_1,e_1),(\lambda_2,e_2),...,(\lambda_p,e_p)$ are the eigenvalue pairs of Σ.

Definition 3: The correlation coefficient index $\rho(Y_k, X_i)$ between the k-th principal component Y_k and the i-th component of the original variable X_i is called the loading of X_i in Y_k.

However, it should be noted that $\rho(Y_k, X_i)$ only measures the contribution of a single variable X_i to the principal component *Y*. When other X_j exists, $\rho(Y_k, X_i)$ does not accurately indicate the importance of X_j to Y. In practice, variables with larger (in absolute value terms) coefficients tend to have larger correlations, so importance measures for univariate and multivariate result to similar result. Therefore, $\rho(Y_k, X_i)$ helps to explain the principal components.

3.2.3 Basic algorithms and steps

The basic algorithms and steps of PCA can be shown below:

(1) Collect the sample workers of the p-dimensional random vector $X = (X_1, X_2, ..., X_p)$, list observation data matrix $X = (x_{ij})_{n \times p}$;

(2) Preprocess the raw data in the sample array converting the original data into positive indicators (the raw data preprocessing was already discussed in Chapter 2), and then use the following formula:

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\mathrm{var}(x_j)}} \qquad i = 1, 2, ..., n \quad j = 1, 2, ..., p \tag{3-6}$$

Where $\overline{x}_j$ and $\sqrt{\mathrm{var}(x_j)}$ are the mean value and the standard variance of the j-th variables. Standardize the resulting data to obtain a standardized array Z.

$$Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_n' \end{bmatrix} \tag{3-7}$$

(3) Calculate the sample correlation coefficient matrix R for the matrix Z:

$$R = \left[r_{ij} \right]_{p \times p} = \frac{Z'Z}{n-1} \tag{3-8}$$

(4) Solve the Eigen-equation of sample correlation coefficient matrix R to obtain p eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$

(5) Obtain the main component $Y_i = u_i' X, i = 1, 2, ..., p$, or $Y = UX$, where

$U = \begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_n' \end{bmatrix}$, $u_{ij} = z_i' b_j^0$, b_j^0 is the unit eigenvector.

The corresponding PCA analysis process algorithm is shown in figure 3.1:

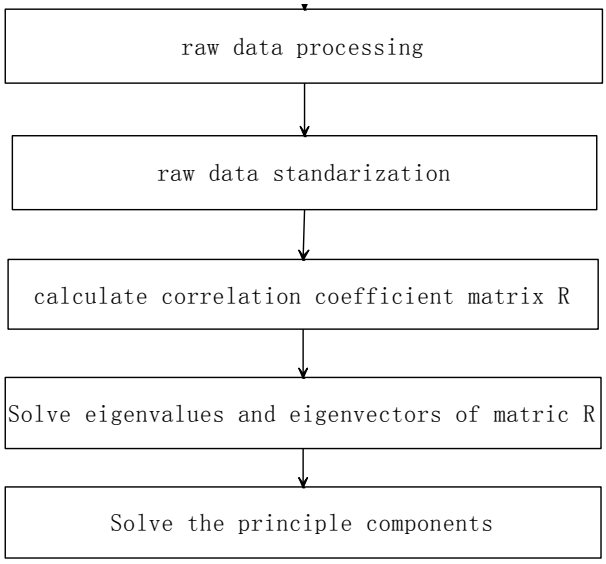


Figure 3.1: The corresponding PCA analysis process algorithm

3.3 Data process

In this section, PCA data processing is discussed.

3.3.1 Data requirements

A basic assumption of the PCA method is that the values of each case corresponding to each component obey the normal distribution.

Usually, especially when the number of cases is large, the assumption is true. However, when the number of cases is small, or the degree of discretization of the criteria values is high, it cannot be assumed that the values of the component also follow the normal distribution. In addition, according to the Law of Large Numbers in mathematical statistics, as the number of evaluation objects increases, the average level and dispersion degree of evaluation indicators tend to be stable, so the covariance matrix also tends to be stable, increasing the accuracy of evaluation results. The PCA is suitable for comprehensive evaluation of large data set. Therefore, in this sense, the PCA method is not applicable to situations where the number of cases is small and the number of components is relatively large.

Similarly, when the PCA is used for multi-index calculation, the normalization (Z-Score) method is often used for dimensionless processing. The premise of standardized application is that the number of data is larger, and it is better to exceed the number of large samples and the sample data is much larger than the index.

On the other hand, the advantage of PCA is that it can turn related data into irrelevant data. In a sense, we also require that the original data have a certain degree of correlation. Otherwise, we lose the significance of applying PCA. In general, the relationship between indicator data is related to the following conditions:

(1) n variables are completely correlated

At this point, the n-1 variables are deleted, and the objects to be evaluated can be sorted. However, this is not really multivariate sorting, and PCA is not useful.

Since the principal components are mutually independent variables, that is, they meet the property 1; many literatures believe that the principal component analysis can eliminate multiple correlations of the original variables. However, in fact, the multiple correlations of variables will distort the real data information from both direction and quantity.

(2) n variables are completely uncorrelated

At this time, these variables cannot be compressed by PCA. Or, usually, the starting point variable correlation matrix of PCA becomes a diagonal matrix, and the de-correlation of PCA does not exist.

(3) There is a certain correlation between n variables

There are certain correlations between n variables, so you can use PCA to compress the data. In this case, the higher the degree of correlation between variables, the better the PCA works.

The relation between variable correlation and feasibility of PCA is shown in the figure 3.2. In real life, most of the indicator variables are not completely correlated, so the PCA is of great use.

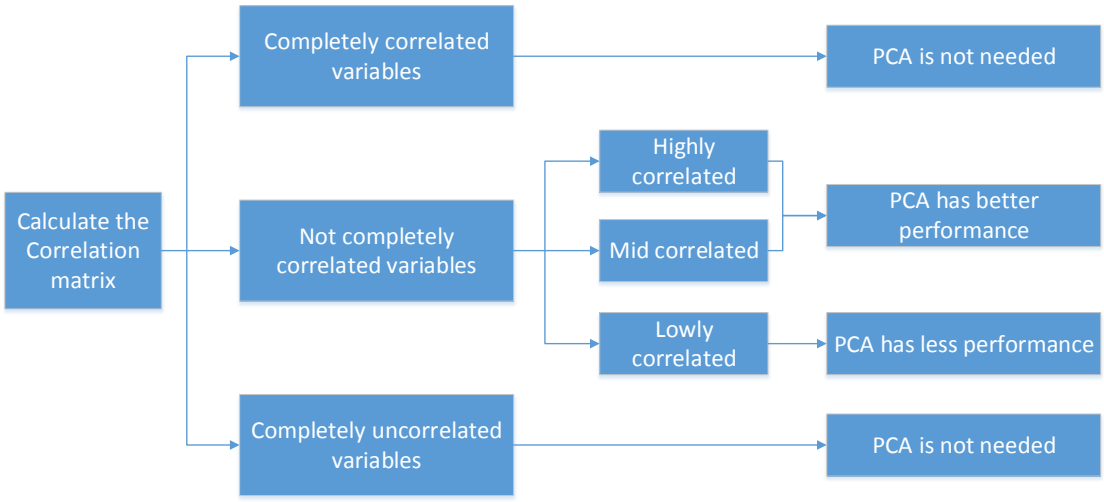


Figure 3.2: The PCA application situations

Summarizing, when performing PCA, the following requirements are imposed on the data:

The number of data required should be large, so that the sample data is much larger than the index. It is not applicable to the case where the number of cases is small and the number of components is relatively large; the variables can have a certain correlation, but in practical applications, it is incorrect to artificially create multiple correlations for variables for comprehensive evaluation. Since most of cases can meet the above requirements in practical applications, therefore the application prospect of PCA is very broad.

3.3.2 Data standardization

The key to principal component analysis is to find the principal component, and its tool is the covariance matrix. Since the covariance matrix is susceptible to the dimension

and magnitude of the indicator, the original data is often dimensionlessly processed. For non-dimensionalization, the standardization method was mostly adopted in the past. We believe that no matter what dimension decreasing method is adopted, there will be information loss because the division in non-dimensionalization is a geometrical similarity transformation, and similarity transformation necessarily changes the data structure. The variable information of the variables is changed, so the sum of the variance before and after the transformation is not equal, which can be proved by mathematics. In general, the original data contains two parts of information: one part is the difference information of each indicator's degree of variation, which is reflected by the variance of each indicator; the other part is the relevant information about each indicator's mutual influence degree, which is derived from the correlation coefficient matrix.

In the chapter 2, five normalization methods have been introduced. The type of standardization that should be adopted depends on the information we need to capture: either the relevant information of each index, or the discrete information of sample points, or the aggregated information of sample points. Whatever a reasonable explanation the calculation results, there is no fixed format. In actual operation, we can use several methods at the same time and decide the trade-off based on actual results.

3.4 Summary

This chapter introduced the basic principles and properties as well as the basic steps PCA by understanding the theoretical basis of principal component analysis, and lists the basic steps of principal component analysis. In addition, two important points of PCA are discussed in the end of the chapter: data requirements and data standardization. The figure 3.3 summarizes the main concepts introduced in this chapter.

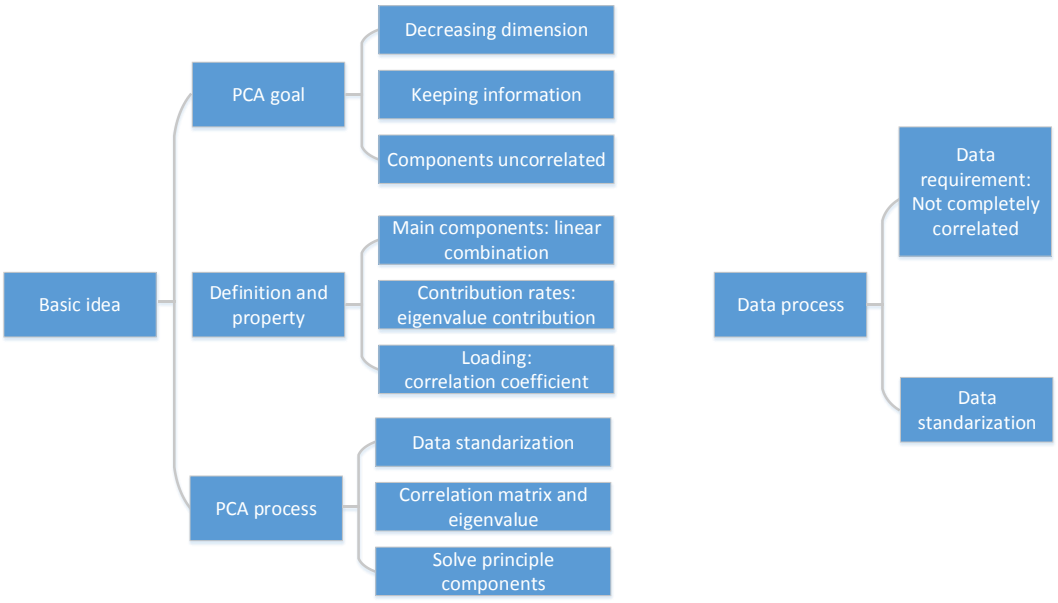


Figure 3.3: PCA concept clarification

4 THE LOAD CLUSTERING MODEL BASED ON PCA

4.1 The load clustering model based on PCA

With the development of demand side response programs and the huge amount data from advanced metering infrastructure (AMI) system, the load profile clustering techniques are applied to classify customers according to their electricity consumption patterns, as well as to evaluate their overall energy consumption trends at a glance.

This paper applies data preparing technology and PCA analysis theory basing on the user load characteristic, so that the user load is divided into the types which have the same or similar load characteristics individually. According to the traditional load profile classification in the chapter 2, we will mainly focus on the active load, customer load, daily load, level-two load

Generally, the clustering of the load profiles using PCA can mainly be divided into five stages:

a. Loading raw load data from the database;

b. Load data preparation;

c. Data simplifying with PCA method;

d. Load profile classification;

e. Applications

The model process is shown in figure 4.1.

The load classification has been made using real load data of UNARETI, the Electricity and Gas Company of the A2A S.p.A. in Milano.

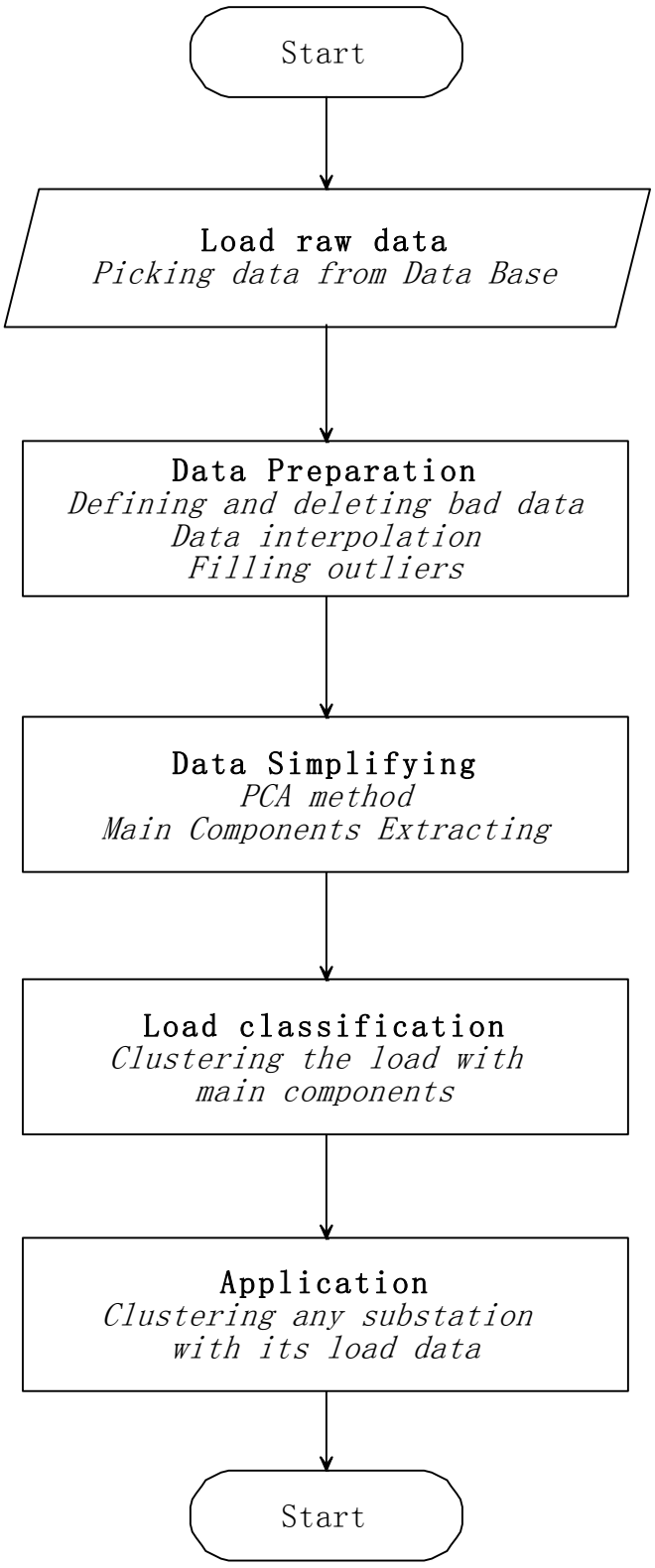


Figure 4.1: The clustering stages of the load profiles

4.2 Loading raw load data from the database

Raw load data used for the analysis are loaded from the database of the company UNARETI in Milano, from nearly 4000 substations located all around Milano. To understand the load data in the database, it is necessary to figure out the transmission and distribution (T&D) system in Italy. Then, it is also necessary to know which information is in the database and pick out the important information to form a new refined database.

4.2.1 T&D system in Italy

The transmission and distribution system (T&D) consists of thousands of transmission and distribution lines, substations, transformers and other equipment scattered over wide geographical area and interconnected so that all function together to deliver power as needed to the utility customer

As a consequence of this hierarchical structure of power flow from power production to energy consumer, a power delivery system can be thought of very conveniently as composed of several district levels of equipment, as illustrated in figure 4.2 shown below:

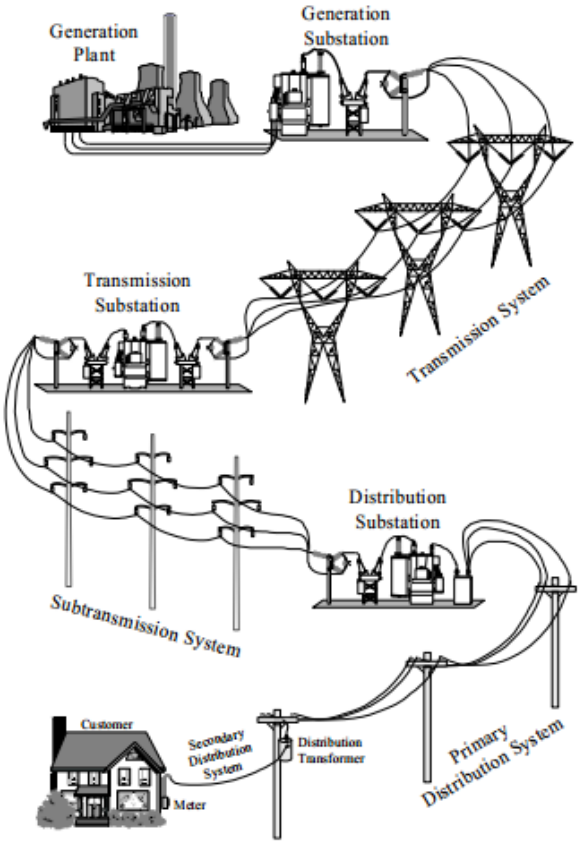


Figure 4.2: The power delivery system

During the transmission and distribution of the electricity, there are mainly 7 levels to consider: the transmission level, the sub-transmission level, the HV/MV substation level, the feeder level, the lateral level, MV/LV substations and the secondary and service level

The transmission level

The transmission system is a meshed network of three-phase lines operating, in Italy, at a voltage generally between 220 kV and 380 kV.

The sub-transmission level

Sub-transmission is that part of the utility system which supplies distribution substations from bulk power sources, such as large transmission substations or generating stations. A typical sub-transmission line may feed power to three or more HV/MV substations, in Italy, at voltage from 132 kV to 150 kV.

The HV/MV substation level

HV/MV substations, the meeting points between the transmission grid and the distribution feeder system, are where a fundamental change takes place within most T&D. Very often a HV/MV substation will have more than one transformer (two is a common number).

The feeder level

Feeder transmits power from HV/MV substations to MV/LV substations. or substation to the distribution points. Feeders operate at the primary distribution voltage. The most common primary distribution voltages in use throughout Italy are 15 kV and 23 kV. An HV/MV substation has between one to forty feeders.

The lateral level

Laterals, short stubs or line segments that branch off the main feeder lines, represent the final primary voltage part of the power journey from MV/LV substation to the customer.

MV/LV substations

MV/LV substations, the meeting points between the feeder system and the distribution secondary circuits, is where transformers lower voltage from primary voltage to the utilization or customer voltage, normally 230/400 Volt.

The secondary and service level

Secondary circuits fed by the service transformers, route power from utilization voltage within very close proximity to the customer.

UNARETI runs the distribution 23kV networks, which is the HV/MV substation level, the feeder level, the lateral level, the MV/LV substations and the secondary and service level.

4.2.2 Database in UNARETI company

After knowing the T&D system in Italy, it is easier to get to understand the database in UNARETI Company. Here we focus on the load in primary side of MV/LV substations, because the MV/LV substations are the stations nearest to the customers. The number of the MV/LV substations is really high (In Milano, the total number of MV/LV substations is around 6000, 4000 among which are equipped with the AMI system). The analysis and classification for the MV/LV can benefit the power distribution company understanding the customer and improve planning. Usually, a MV/LV substation has one or two transformers, called TR1 and TR2.

In the database of UNARETI Company, there are several types of information about MV/LV substations:

1. Daily load profile

The daily load profile is one of the most important data we should focus on. Table 4.1 shows the original daily load profile data and the explanation.

Table 4.1 The original daily load profile database

Data	Explanation	Example
IDRTU	Substation ID number	1
CodificaCliente	Substation code	E02073

Location	Substation location	Piazza Tirana 32
DATAORA	Time of the recorded data	01/12/2016 0.00.00
TMED	Average temperature	19.06
TMAX	Maximum temperature	19.42
OMAX	Recording time of maximum temperature	01/12/2016 0.00.00
TR1MAX	Maximum load of transformer-1	252.3839797
OTR1MAX	Recording time of TR1 load	01/12/2016 13.30.00
TR2MAX	Maximum load of transformer-2 in	
OTR2MAX	Recording time of TR2 load	

It can be seen that there are 11 data available in daily load profile, which is related to the daily load, temperature, the time and the geographic information.

2. Coordinates

The coordinates information shows the detail geographic information for each substation. With coordinates information, the substation should be located and named exactly. Table 4.2 shows the coordinates information data and the explanation.

Table 4.2 The original coordinates database

Data	Explanation	Example
CodificaCliente	Substation code	A01004
x	x distance from Rome origin(m)	1511569
y	y distance from Rome origin(m)	5037759

latitude	Latitude of the substation	45.49
longitude	Longitude of the substation	9.14
V	Voltage level(V)	23000

In the coordinate's information of the substation, there are two ways to record the geographic information of the substations. Considering the convenience, the latitude and longitude information is more popular.

3. LV customer information

Every substation supplies a specific number of LV customers. The information of LV customers is important in classifying the load.

Table 4.3 shows the LV customer information, with the data and the explanation.

Table 4.3 The original LV customer information database

Data	Explanation	Example
CodificaCliente	Substation code	A01001
LV Num	Total number of LV customers	199
Total power	Total contractual power	919.95
RATIO	Ratio of power and customer number	4.62

It is important to point out the definition of the data RATIO:

$$RATIO = \frac{Total\ contractual\ power}{LV\ customers\ number} \tag{4-1}$$

Ratio is the power consumed per LV customer unit in the distribution area. It can represent the type of the customers supplied by the MV/HV substations.

4. Overloads

In the database, there are is information about overloads. Table 4.4 shows the overload time during one year.

Table 4.4 The original overload number information database

Data	Explanation	Example
IdRtu	Substation ID number	3
Indirizzo	Substation address	Via Padova 115
CodiceAem	Substation code	E04439
gennaio	Overload times in Janary	0
febbraio	Overload times in Feberary	0
marzo	Overload times in March	0
aprile	Overload times in April	0
maggio	Overload times in May	0
giugno	Overload times in June	4
luglio	Overload times in July	2
agosto	Overload times in August	0
settembre	Overload times in September	0
ottobre	Overload times in October	
novembre	Overload times in November	
dicembre	Overload times in December	
Totale	Total overload times in one year	6

The overload information can represent two kinds of information in total: the reliability of the substation and the month peak load.

5. Others

In the database of UNARETI Company, there is other information we can obtain. For instance, there is the fan data, which shows the fan information in the transformers everyday. And there is the information of the daily weather and temperature and so on. All the data is stored in the database in the format of '*.csv'.

4.2.3 Refined database building

The database from the company is various and complete. But for the calculation, there is lots of weakness:

1. There are many same data in different formats;

2. Some data has the same function;

2. The data is in different files, which leads to the difficulty to deal with;

3. There are lots of useless data in the exact case in the file, which occupies too much space;

4. The names and the symbol of the data are not uniform;

5. The data is not clearly classified;

6. The data is not targeted

So here we only use some of the data. So we try to unite the data and refine the database according to our case. Table below is the refined data for the exact issue.

Table 4.5 The refined database

Data type	Symbol	Explanation
Substation information	Id	Substation ID number
	Code	Substation code
Daily load	TR1MAX	Maximum load of transformer-1

	OTR1MAX	The time to of the peak daily load of TR1
	TR2MAX	The maximum load of transformer-2 in one day
	OTR2MAX	Time to obtain the maximum TR2 load
Customer information	LV Num	Total number of LV customers
	RATIO	Ratio of power and customer number
geographic information	latitude	Latitude of the substation
	longitude	Longitude of the substation

Finally, the daily load profile in the year 2016 is loaded for all the nearly 4000 substations, each of which has 366 daily data because that 2016 is a leap year.

4.3 Load data preparation

The load data preparation usually includes defining and deleting bad data, data interpolation and filling outliers.

4.3.1 Defining and deleting bad data

For the load profile, there are only 4000 out of 6000 MV/LV substations monitored in the system. Among those 4000 substations, there are many of unpleasant data.

Unpleasant data includes missing values, 0 values, and straight line load values. Unpleasant data can be classified into two types: fixable data and unfixable data. The

unfixable data can be called bad data, mainly including both 0 values and long series of missing data. The causes of the bad data in the AMI system are probably:

1. The measuring machine faults;

2. The manual mistakes;

3. The system updating and repairing;

4. The overdue of the AMI system installing;

5. The substation repairing

The only way to deal with bad data is to detect it and delete or recover.

The physical identification is usually used in bad load data detection. As to the daily load data, it is important to see if it is fixable or not. There are two index to identify the bad data:

1. The total missing data

The total missing data refer to the total missing data compared to the data set. It is a large scale index, which refers to the overall goodness of the data set.

2. The maximum continuous missing data

The maximum continuous missing data refer to maximum number of adjacent missing data. It is a partial index, which refers to the overall goodness of the data set. Too much continuous data missing will cause a large error when trying to fix it.

Thus with the first index, all the substations which have too few data (less than 200 daily load data) in total are deleted.

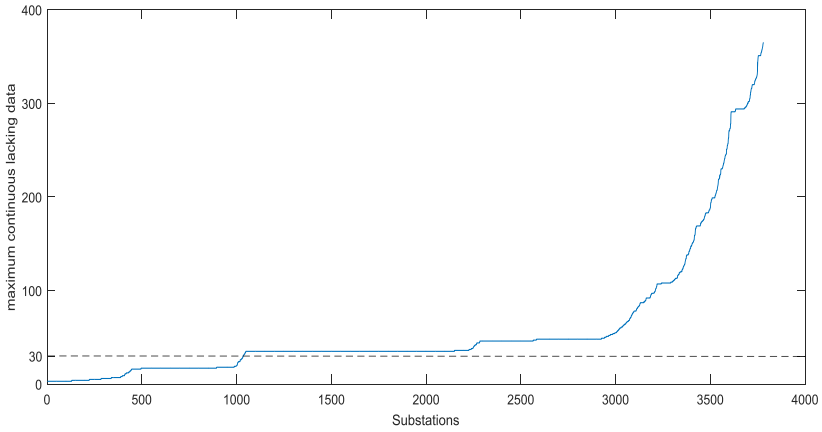


Figure 4.3: The continuous lacking substations

As to the second index, figure 4.3 shows the maximum continuous missing data for all the substations. From the maximum gap of missing data, for the accurate information of the data and the quantity of data reserved, it is better only to use the substations whose gap is less than 30, that is, one month. With this method, there are still the data of 1036 substations remained to be analyzed.

4.3.2 Data normalization

The user load data obtained through the power system load measurement device has a large difference in the value range, and these differences will have a great impact on the classification result. Therefore, sample data should be normalized before classification to eliminate the influence of these differences.

In chapter 2, five normalization methods were introduced.

In order to classify the substations with the daily load value, the load value should be normalized to decrease the exact load value influence. There are two points we should focus on:

1) The shape of the load should be maximally preserved;

2) The peak load of all substations should be standardized to be the same to avoid the load value influence.

In table 4.6 there is the comparison of the five methods:

Table 4.6 The typical interpolation methods comparison

Name	Weakness	Strength
Feature scaling	The peak value is not fixed after normalization	The shape is reserved
Feature scaling standardization	The peak value is not fixed after normalization; The shape is changed	
Student's t-statistic standardization	The peak value is not fixed after normalization; The shape is changed	
Fractional scaling normalization	The peak value is not fixed after normalization;	The shape is reserved
Max load normalization		The peak value is 1 for all the substations after normalization; The shape is reserved

So from the analysis above, the max load normalization is the most proper normalization method in this case.

4.3.3 Data interpolation

During the load collection process, how to deal with the fixable unpleasant data should be considered. Interpolation is an effective method in dealing with missing values of non-bad data. In the mathematical field of numerical analysis, interpolation is a method of constructing new data points within the range of a discrete set of known data points.

In the Chapter 2, five interpolation methods are introduced. According to the features of all the interpolation methods, the PCHIP+ARMA method is proper in this case. For instance, figure 4.4 shows the real profile of one of the MV/LV substation. It shows the

original load profile with some missing data, the load profile after the PCHIP interpolation and the load profile after PCHIP+ARMA method.

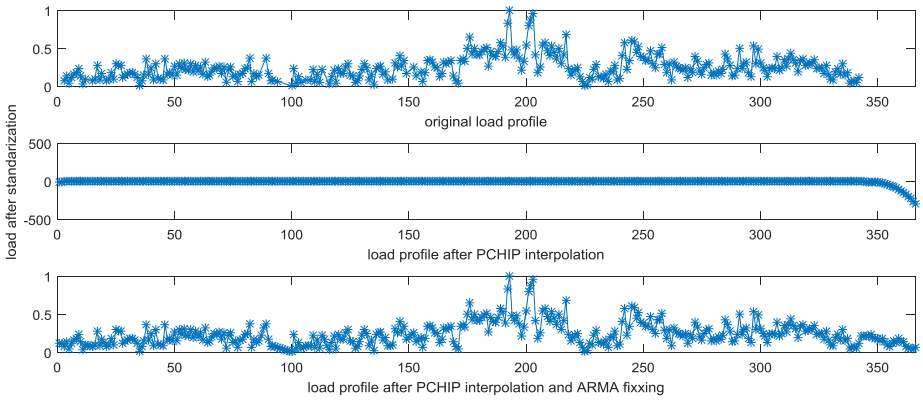


Figure 4.4: The PCHIP method compared with PCHIP+ARMA method

Thus PCHIP is good to fix the lacking data, and ARMA to fix the boundary. Putting two methods together can achieve a better data fixing.

4.3.4 Outliers processing

In statistics, an outlier is an observation point that is distant from other observations.

The data from the company has some outliers due to machine mistakes. The database in UNARETI is really large, so some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected.

In this case, 3-σ rule can be used in the outlier detection in daily load, considering that 0.2% possibility is too little to be reality.

The daily load can be signed as $X = (x_1, x_2, ..., x_n)$ with the mean value $\overline{X}$ and the standard deviation σ. Then if $x_i > \overline{X} + 3\sigma$ or $x_i < \overline{X} - 3\sigma$, we can say that is an outlier.

After pointing out the outlier, due to the dispersion of the outlier, the PCHIP method is used to fill gaps or outliers.

The figure 4.5 shows the process of filling the outliers:

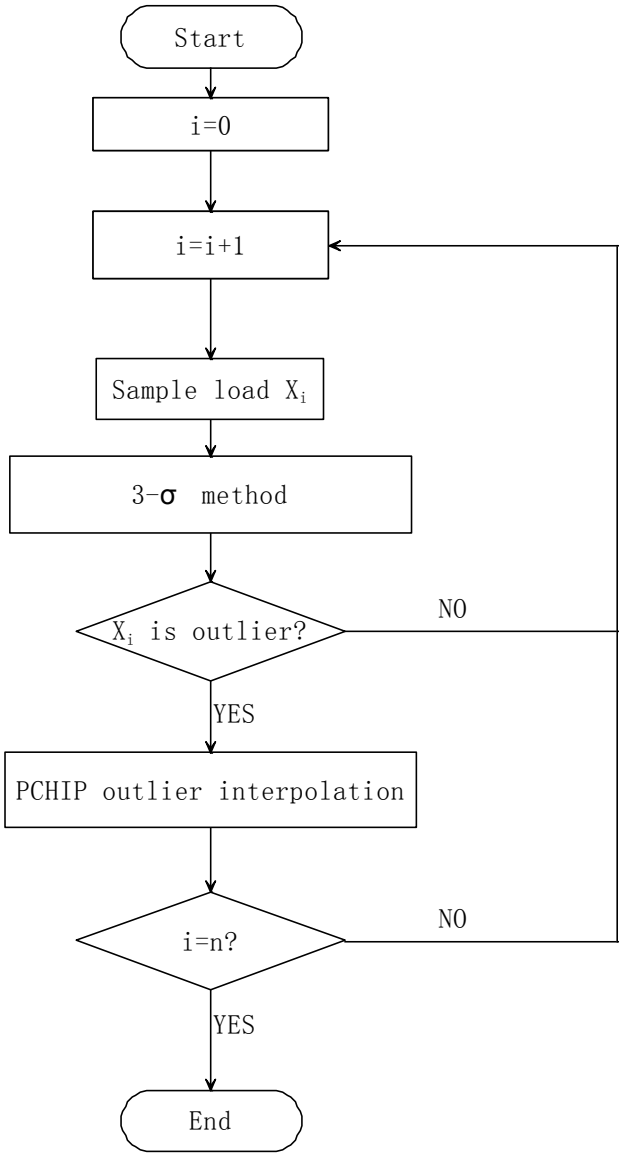


Figure 4.5: The process of filling the outliers

The figure 4.6 shows an example of outlier in the load profile of one of the MV/LV substations and how it has been fixed. It can be seen that the two peak load are outstanding according to the data near and then they are probably outliers. From the result, several outliers are detected and filled.

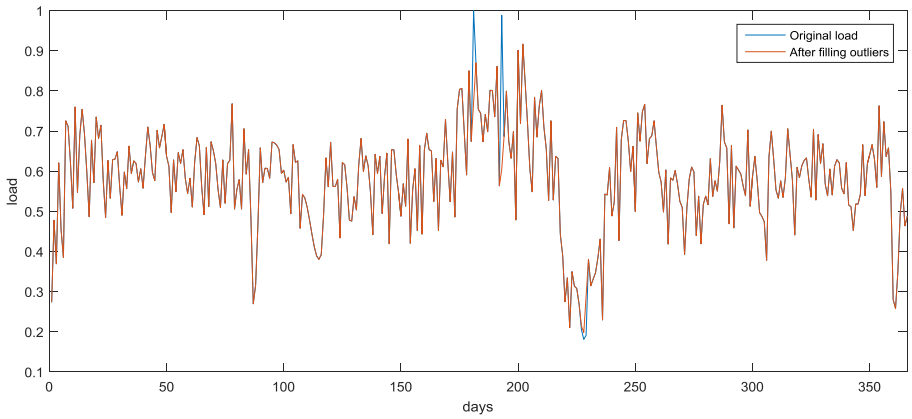


Figure 4.6: Example of an outlier fixing

4.4 PCA method in load profile

4.4.1 PCA in load profile

After building the database, it is easy to see the dimension of the database is large and complex. Table 4.7 shows the detailed database scale:

Table 4.7 The detail database scale

	Each substation data size (x double)	All 4000 substations data size (x double)
ID	1	4000
geography	2	8000
Daily load	732	2928000
customer	2	8000
in total	737	2948000

Table 4.7 shows that for each substation, there are at least 737 data with different meanings. To analyze the 732 daily load data to classify the substations is difficult. Therefore, PCA method is used here to reduce the data size.

4.4.2 The process to apply PCA

We apply the already normalized load profile for each substation, with the variables to the daily load in the year of 2016. Figure 4.7 shows the process of PCA in load profile analysis:

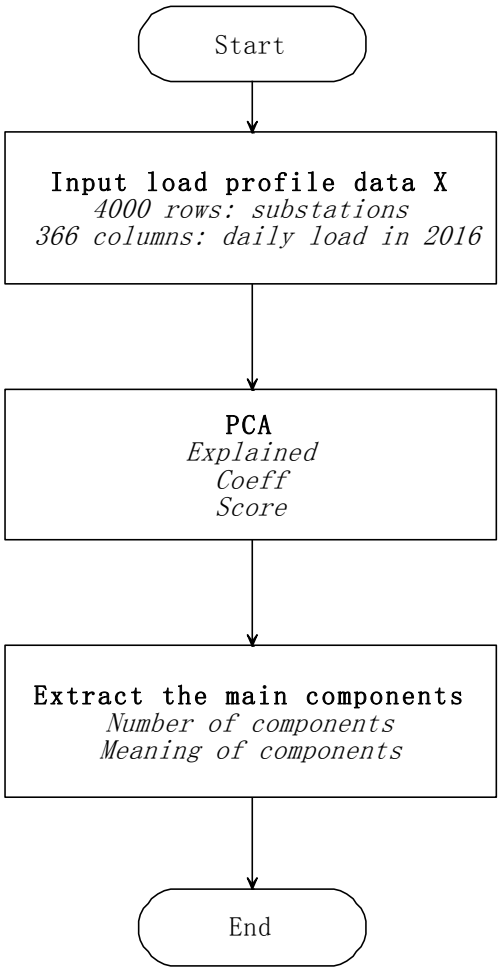


Figure 4.7: The process of PCA in load profile analysis

1. Input load profile data

PCA requires to input a data matrix X, where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of feature (say, the results from a particular sensor).

Thus, after the data processing, the daily load matrix should have 1036 rows representing substations and 366 columns representing daily load in 2016 for each substation.

2. PCA

There are three important output in the matlab, which are explained, coeff and score.

1) Explained: which represents how much percentage of the whole information the component can explain.

2) Coeff: it represents the loadings of the components, which are the coefficients on the daily load to calculate the component.

3) Score: it represents the scores of the substation, which are the values of the components calculated for the substation.

3. Extract the main components

To extract the main components, two are the important indexes:

1) The number of main components: It depends on the total percentage of the total information the components can explain. The goal is to explain the most of the information with the least number of the components.

2) The meaning of the main components: usually, every component has a meaning. It can be concluded from the coefficient from the PCA. For example, the positive coefficients refer to the variable is proportional to the components. And the more the coefficient it is, the more contribution the variable gives.

4.5 Load profile classification and application

After applying the PCA, the dimension of the load data for each substation is decreased from 366 items of data to 2 or 3 main components. In this way, it is easier to classify the substations with only 2 or 3 data each.

For example, when there are 3 main components, a way is to classify the load profile using eight octants. In paper to be sure that different classification has similar quantity of data, it is necessary to choose a proper origin. For this purpose we can use the median point for three components as an origin. Thus the classification model can be shown in figure 4.8:

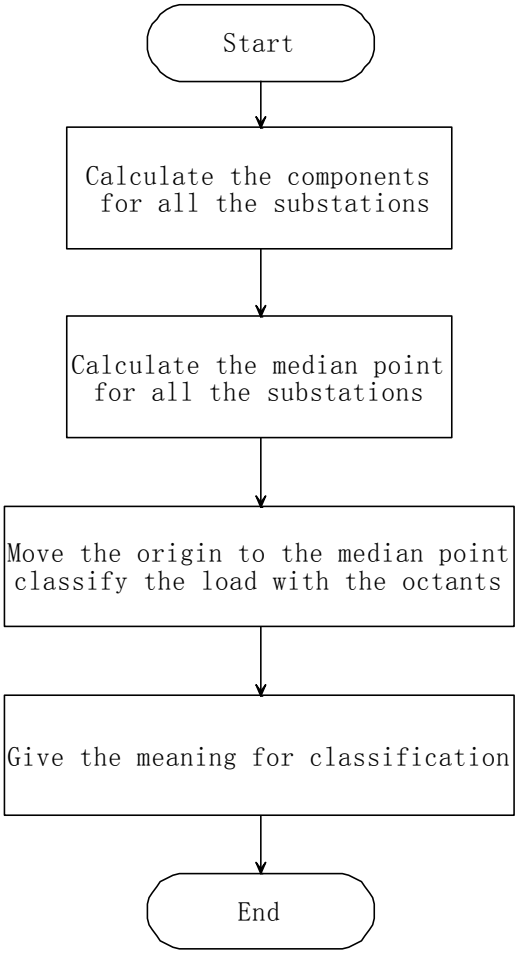


Figure 4.8: The classification model

When a new substation load is known, we just need to calculate the main components. And comparing with the median point, the classification can be obtained by the way that the value is more than the median point or less than median point.

4.6 Summary

In this chapter, the load classifying model based on the PCA method is put forward, including five steps: loading raw load data from the database, load data preparation, data simplifying with PCA method, load profile classification and result applications. Based on the real load data from the distribution company in Milano, a procedure has been developed.

In the step of loading data from the database, the method of refining and simplying the database is introduced and at the end a clean database is built from the raw data. In the

step of load data preparing, the methods of data preparing is applied in the problem, in four steps. First, defining and deleting the bad data with two indexes: the total missing data and the maximum continuous missing data. Second, normalizing data with the max load normalization after the comparison. Third, the PCHIP+ARMA method is used in the data interpolation. At least, 3-σ rule is applied in outlier detecting.

Then, in the next step, the opportunity of applying PCA, and the processing to apply PCA in the load classification are introduced.

Finally, load profile classification model is built, with calculating the components and using the octants to classify MV/LV substations load profile.

5 RESULT OF THE PCA ON LOAD PROFILES

5.1 Analysis of PCA components

5.1.1 Input and output of PCA

The input and output of PCA should be clarified in the beginning, including the data type, dimension and explanations.

The input of PCA is the refined data set modified from the original load data set, with columns corresponding to 366 days of the year, and 1036 rows of selected substations.

The main output of PCA are the explanation matrix, the coefficient matrix and the score matrix.

The figure 5.1 shows in detail the meaning of the input and output

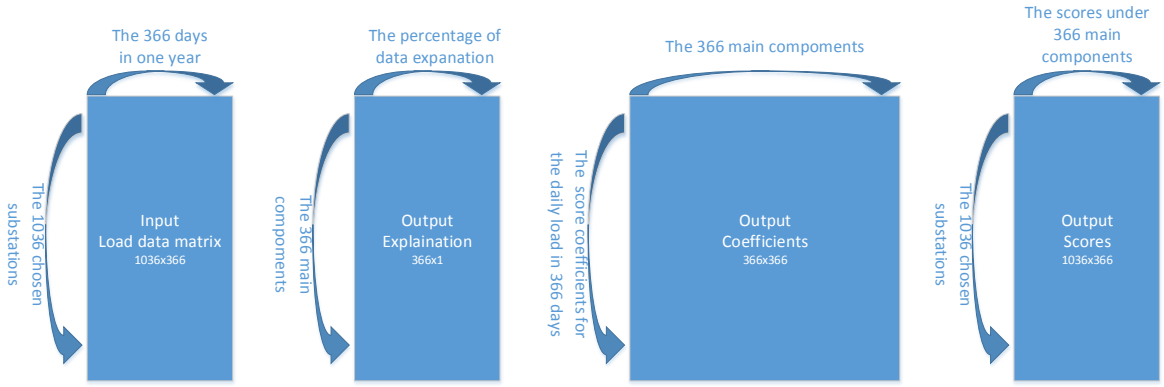


Figure 5.1: The coefficients of the first main component

5.1.2 PCA main components

After applying PCA into the case in Milano, the main components can be identified.

Table below shows the explain matrix of the daily load.

Table 5.1 The explain matrix of the daily load

Main components	Explained(%)	Accumulate(%)
1	51.00	51.00
2	11.98	62.97
3	5.11	68.09
4	3.48	71.57
5	2.46	74.04
6	2.04	76.08
7	1.56	77.64

8	1.32	78.96

From the table it is easy to see that first 3 components can explain almost 70% of the information. Thus 3 main components in the daily load can be extracted. The three components are analyzed one by one.

1. First component: activities load factor

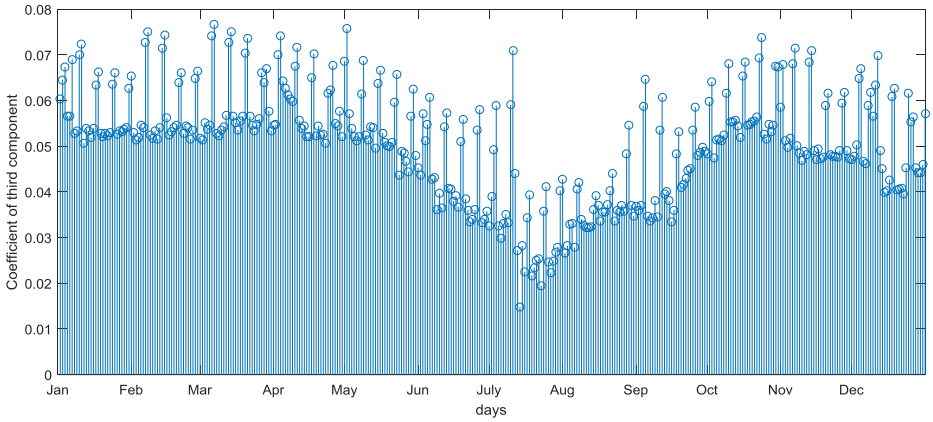


Figure 5.2: Coefficients of the first main component

The figure 5.2 shows the coefficient of the first main component. The coefficients of months out of July and August (When Italians often go to vacations) are similar. And the coefficients in July and August are relatively smaller. So from the coefficients, the average load in months out of July and August is more important and contribute more to the score. In this way, it may represent the activities factor, which means the customers in the substation go on vacation during summer or not. So it may be no-summer load factor.

2. Second component: air conditioner factor

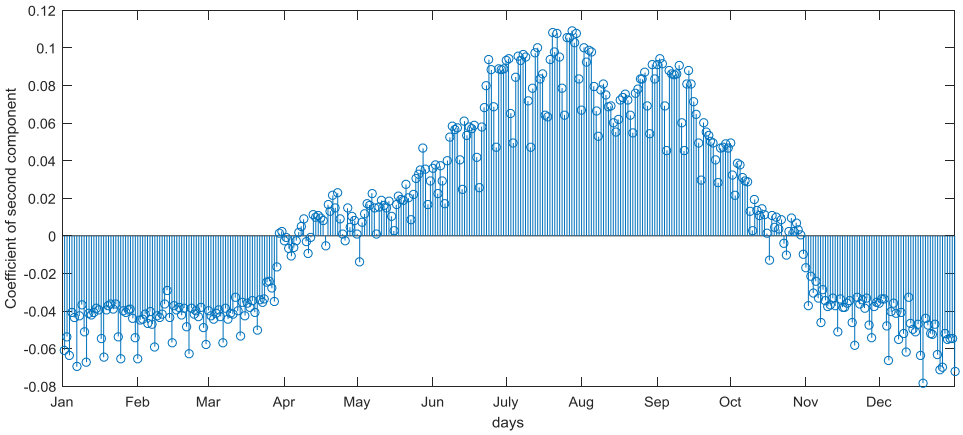


Figure 5.3: The coefficients of the second main component

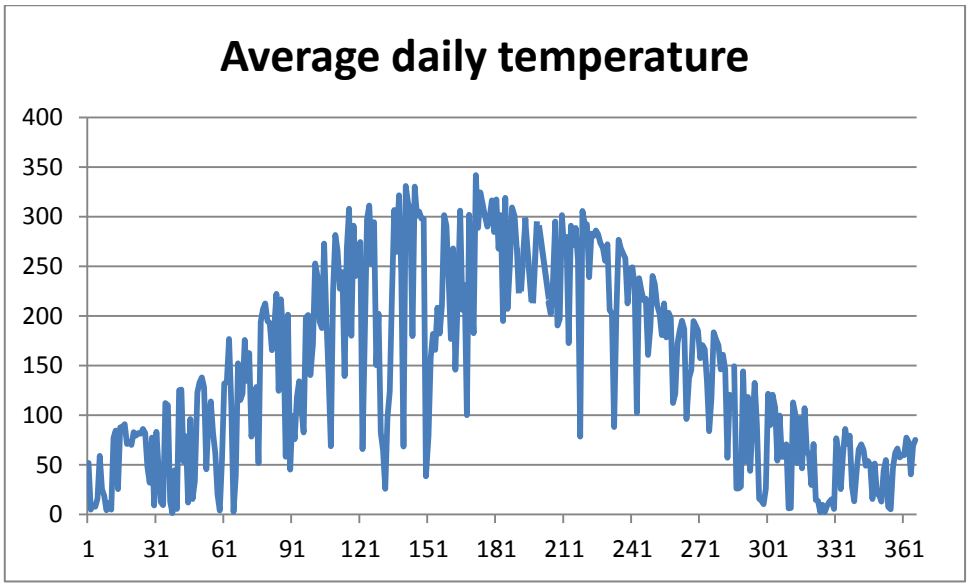


Figure 5.4: Average daily temperature in Milan

The figure 5.4 shows the coefficient of the second main component. Compared with the average temperature in Milano, it is easier to see a similar trend. Thus the second component may be considered as the air conditioner factor, representing the use of air conditioner by customers.

3. Third component: holiday factor

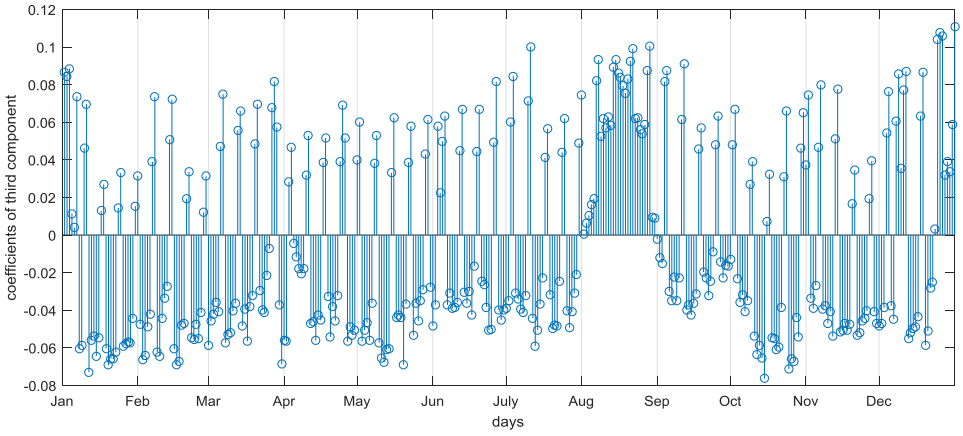


Figure 5.5: The coefficients of the third main component

The figure 5.5 shows the coefficient of the third main component. It can be easily seen that the coefficients of the load during weekends, Christmas, summer or vacations, are

positive while the coefficients of the load in working days are negative. For example, focusing on the first week of February, the second of February in 2016 is Monday, so that component is negative. From the Figure 5.6, it can be easily seen that on the work days, the coefficients are negative and on the holidays (weekend) the coefficients are positive.

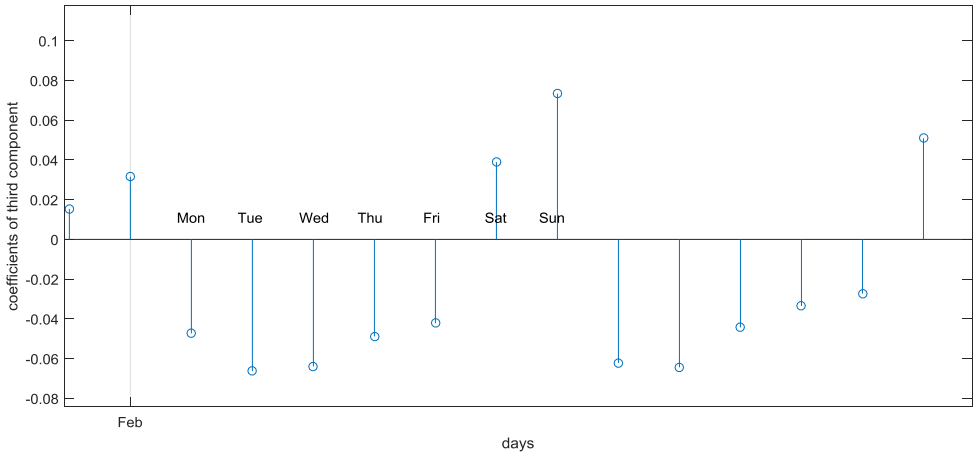


Figure 5.6: Coefficients of the third main component in February

It can be also seen that in Christmas from December 22nd to January 7th the coefficients are positive.

From the analysis above, the third component might be the vacation factor.

5.1.3 The load properties of main components

To prove the hypothesis of the previous section, for each component, the load profile with higher component scores and the load profile with lower component scores are chosen to compare. From the difference of the load profiles, the properties that the main components represent can be probably defined.

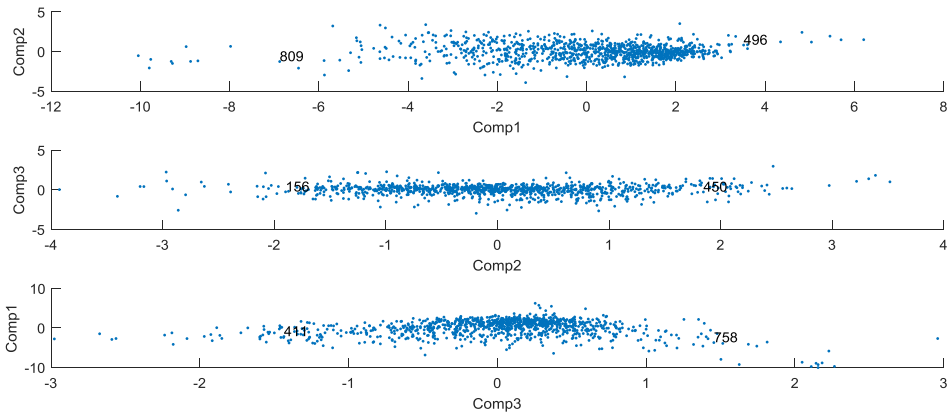


Figure 5.7: The score of all the substations in couple

The figure 5.7 shows the score of all the substations shown in every two components. In the figure 5.7, different couples with different values of three components are chosen to see the load profile. For example, No.809 substation with a lower component 1 score is chosen to compare the load profile with No.496 substation with a higher component 1 score to see the property of component 1. So are No.156 substation and No.450 substation chosen to see the property of component 2. So are No.411 substation and No.758 substation chosen to see the property of component 3.

1. First component

For the first component, the number 809 substation with lower score and number 496 with higher score substation are picked. The figure 5.8 shows the load profile for the two substations:

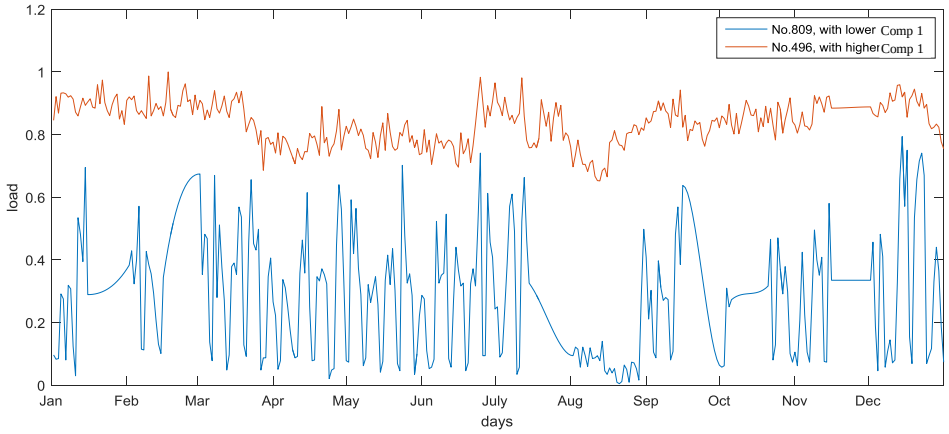


Figure 5.8: Load profile of No.809 and No.496 substations

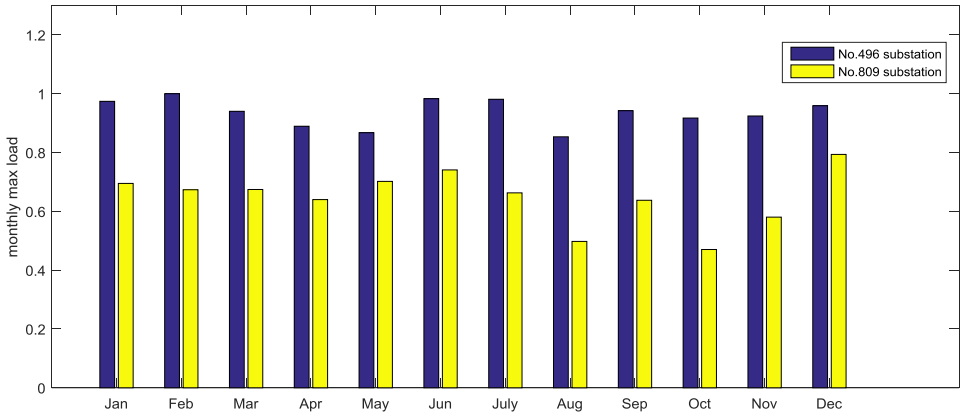


Figure 5.9: The monthly max load of No.809 and No.496 substations

From the monthly max load profile, we can see that for substations with lower first component, the load is less uniform. For example, there is the difference between the summer (August) and all the other months. The substation with higher first component has more uniform load, which may represent the uniform activities in one year. So the assumption seems to be proved.

2. Second component

For the second component, the substation number 156 with lower score and substation number 450 with higher score are picked. The figure 5.10 shows the load profile for the two substations:

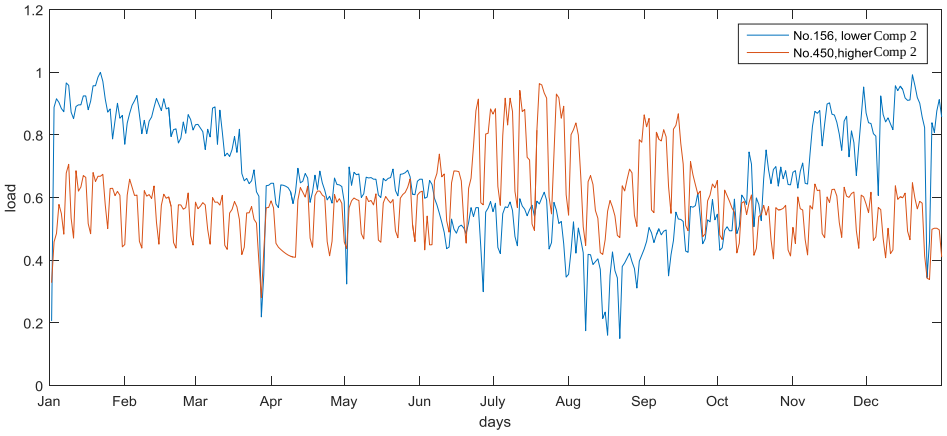


Figure 5.10: The load profile of No.156 and No.450 substations

From the load profile, probably for the substation with higher second component, summer load is higher, which may represent the use of air conditioners. Thus the assumption seems to be proved.

3. Third component

For the second component, substation number 411 with lower score and substation number 758 with higher score are picked. The figure 5.11 shows the load profile for the two substations:

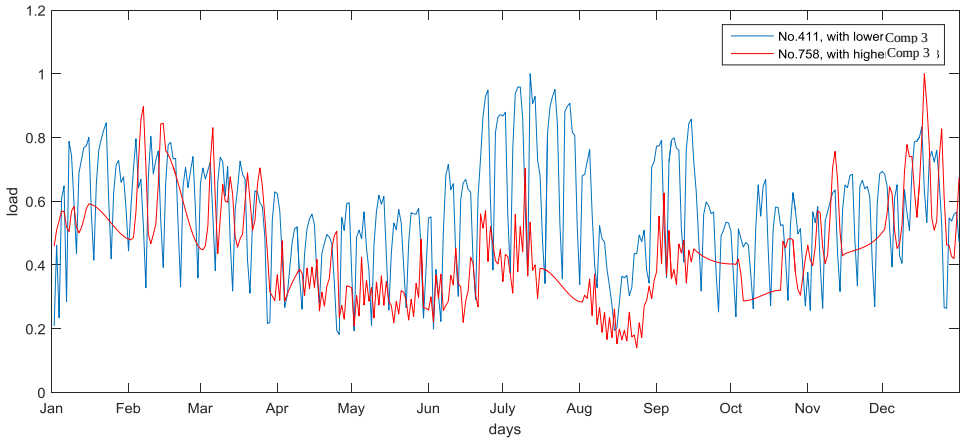


Figure 5.11: The load profile of No.411 and No.758 substations

To see the difference, figure 5.12 shows a time period of a week.

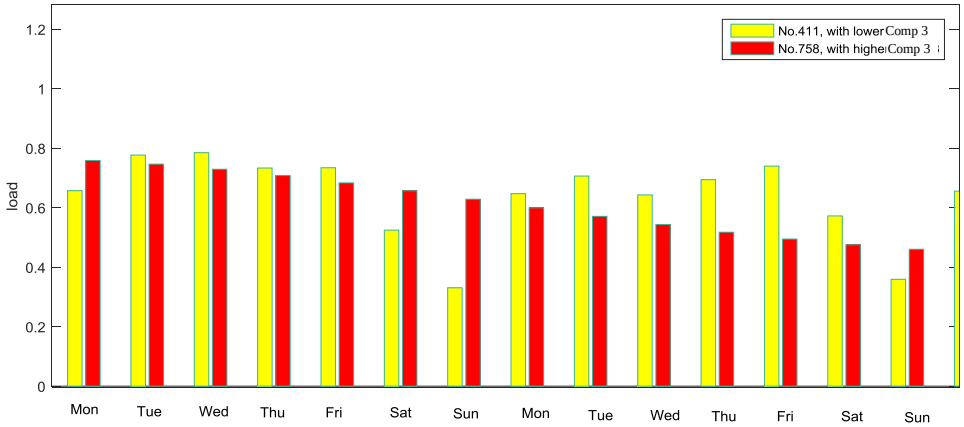


Figure 5.12: The two week load profile of No.411 and No.758 substations

Figure 5.12 shows the load of first two weeks in May. From the load profile, it is shown that No.411 substation has a more working days and holidays difference. For the substation with higher score, the working days and holidays difference is less. Thus the assumption may be proved.

5.1.4 Brief summary

From the analysis above, we can conclude:

The first factor may mean the load difference between summer and the other months, that is if the customers in the substation take the summer vacation or not. So we can call it activities load factor. The larger activities load factor the substation has, the more less differences between August and other months it may have, which means the more probability that the customers have the summer vacation in Italy.

The second factor should be related to the temperature, might be considered as the quantity of air conditioners in use. So we can call it air conditioner factor. The larger air conditioner factor the substation has, the more energy is consumed in summer than in winter.

The third factor should be related to the holidays, might be considered as the difference between working days and holidays. So we can call it holiday factor. The larger holiday factor the substation has, the less differences it has between holidays and work days.

5.2 Result of load profile classification

5.2.1 Classification result

After giving the meaning of each main components, we are able to classify the substations by dividing the value of the three components into eight octants.

The scatter graph of three components for all the substations can be shown below:

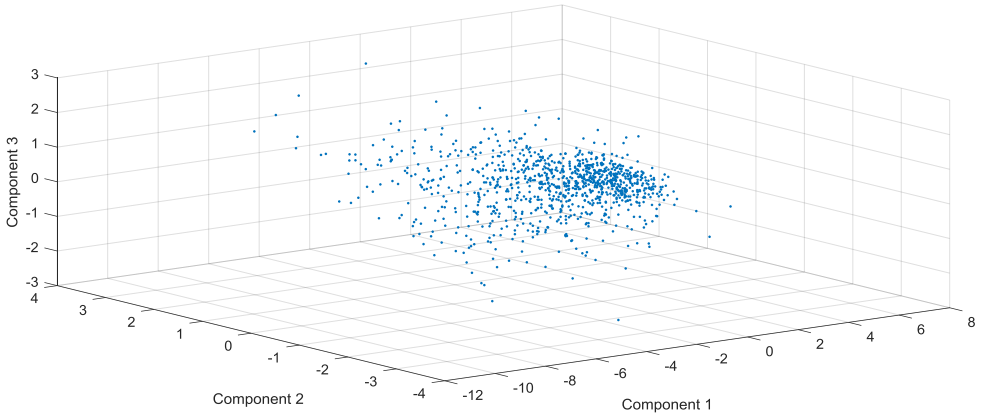


Figure 5.13: Scatter graph of three components

To classify the substations, it is important to ensure the origin.

Thus, we use the medians of three components scores , that is (0.4393, -0.0302, 0.0753), to be the origin.

The figure 5.14 shows the classification method while the table 5.2 shows the meaning of any load type:.

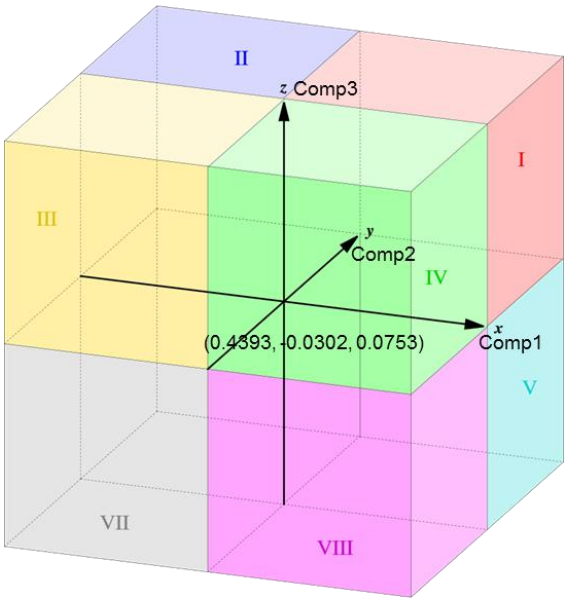


Figure 5.14: Octants classification method

Table 5.2 The meaning of clustering

Cluster	Definition	Main	Load type

	（Comp1,Comp2,Comp3）	characteristics	
Ⅰ	（+,+,+）	Uniform activities, high airconditioner occupancy, low holiday differences	Household out of city center
Ⅱ	（-,+,+）	Non-uniform activities, high airconditioner occupancy, low holiday differences	Household in the city center
Ⅶ	（-,-,-）	Non-uniform activities, low airconditioner occupancy, high holiday differences	Factory or commercial buildings(Non-uniform activity)
Ⅷ	（+,-,-）	Uniform activities, low airconditioner occupancy, high holiday differences	Factory or commercial buildings(Uniform activity)
Ⅲ	（-,-,+）	Non-uniform activities, low airconditioner occupancy, low holiday differences	
Ⅳ	（+,-,+）	Uniform activities, low airconditioner occupancy, low holiday differences	
Ⅴ	（+,+,-）	Uniform activities, high airconditioner occupancy, high holiday differences	
Ⅵ	（-,+,-）	Non-uniform activities, high airconditioner occupancy, high holiday differences	

In the table 5.2, the sign refers to the median value. For example, the symbol (+,-,+) means the type of substation with the component 1 higher than the median value, component 2 lower than the median value, component 3 higher than the median value.

5.2.2 Geographic proof

Actually, not all the types are common in the daily load. Figure below shows the distribution of all the type substations in Milan. It is obvious that there are only 6 types substations in Milan, and type 3 and type 4 still didn’t exist in Milano. What’s more, the type 8, type 7, type 2 and type 1 are most common substations.

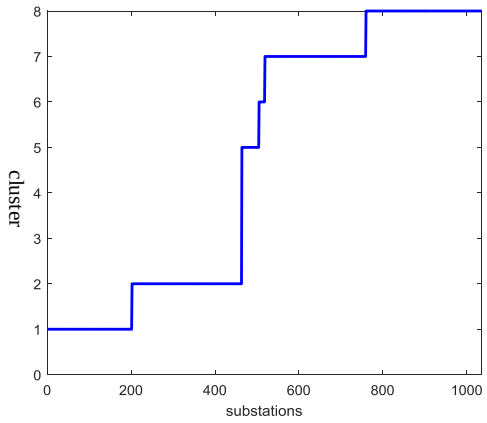


Figure 5.15: Load classifications distributions

Combined with the geographic information, with Google Map, we can sign the eight different classifications into different colors and show it on the map, as shown in the figure below:

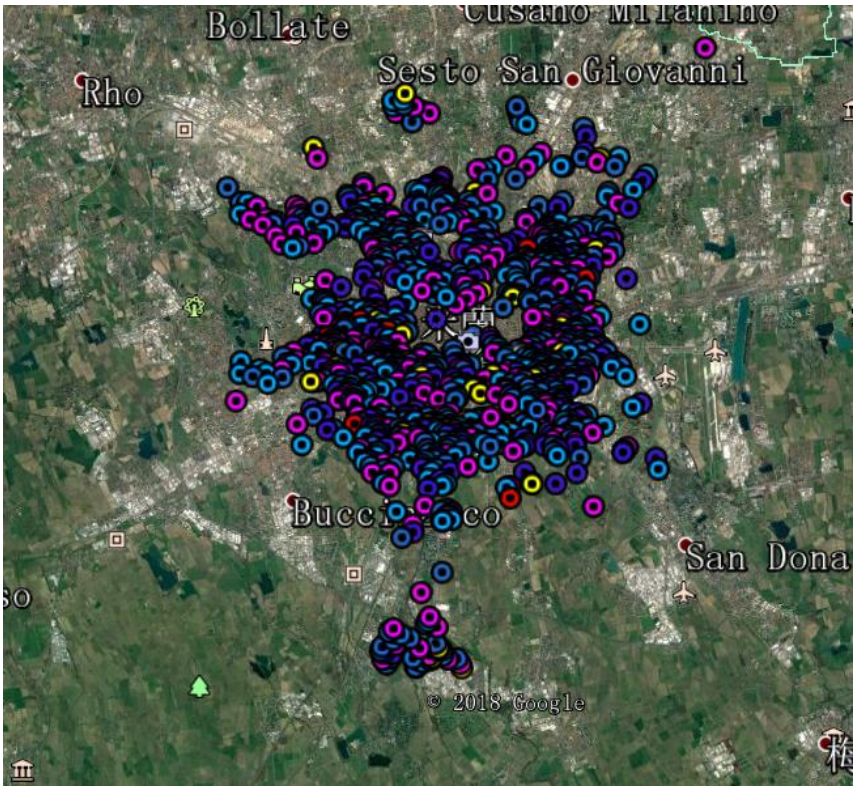


Figure 5.16: Different load classifications in in google map

On the map, some substations can show the properties of the load.

Type 1 is uniform activities, high airconditioner occupancy, low holiday differences, which seems meet what we assumed. It may be household out of city center. For instance, the no.820 substation shown in figure 5.17 is the house out of the city center. And the no.1227 substation is located in the household in Milano Villapizzone, which is far from the city center.

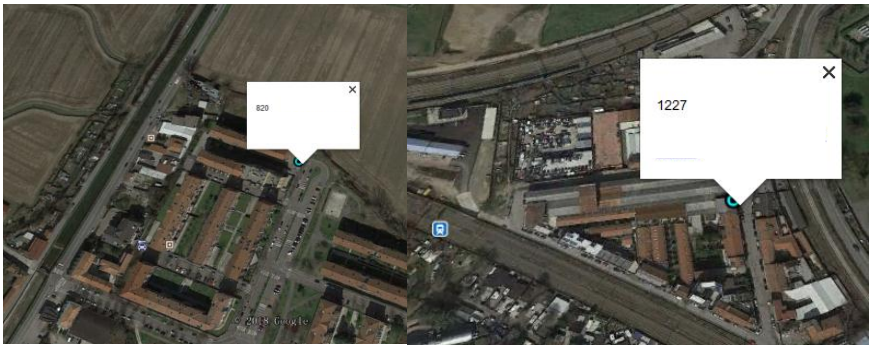


Figure 5.17: Type 1 substation in google map

The type 2 has the character of non-uniform activities, high airconditioner occupancy, low holiday differences, which may be household in the city center. For instance, the no.227 substation shown in figure 5.18 is near city center Duomo, and the no.1393 substation is near city center Lanza. They seem to be the household substations, which meets our assumption.

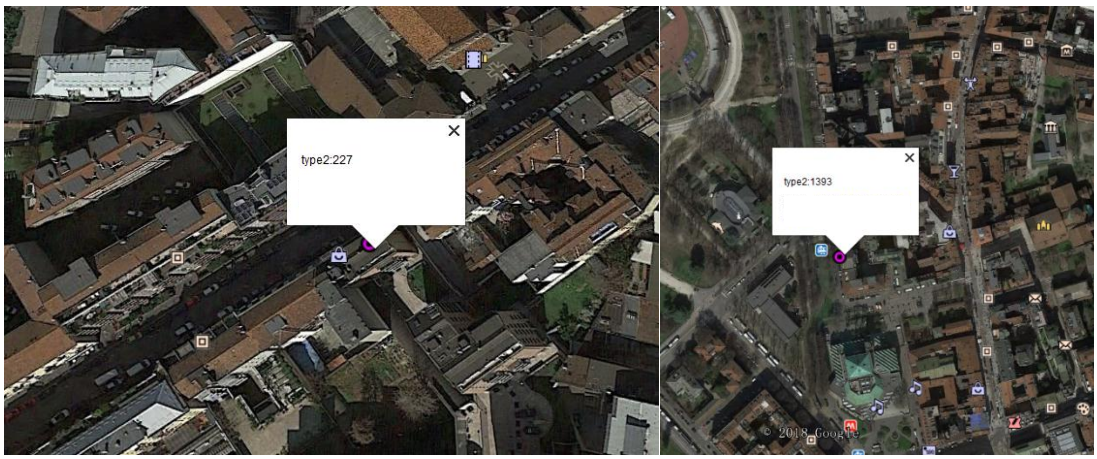


Figure 5.18: Type 2 substation in google map

Type 7 has the character of non-uniform activities, low airconditioner occupancy, high holiday differences. It can be factory or commercial buildings with non-uniform activities. For instance, the no.93 substation shown in the figure 5.19 is for the milano

hotel. There are differences between summer vacation and other months in one year. The no.858 substation is the residence of university Bocconi, which is also a non-uniform commercial building.

Figure 5.19: Type 7 substation in google map

The type 8 has the characters of uniform activities, low airconditioner occupancy, high holiday differences. It might be factory or commercial buildings with non-uniform activities. For instance, the no.1062 substation shown in the figure 5.20 is in the commercial center of Milan, and it is probably for the bars and restaurants, which are uniform load. And no.419 substation is for the doctors and shops. So our assumptions are probably proved.

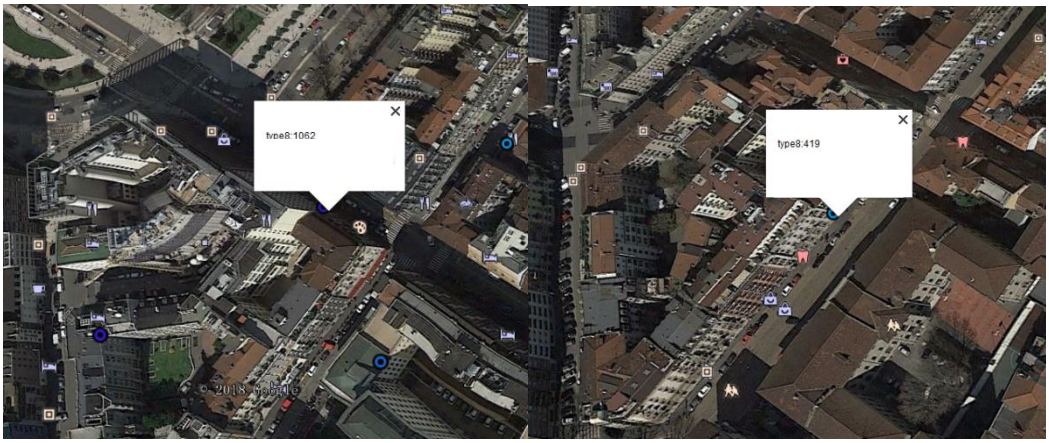


Figure 5.20: Type 8 substation in google map

5.3 Load profile classification by the daily peak load

For planning purpose, it is also important to know the maximum load day of the substations.

Considering the data of all the 1036 MV/LV substations, the histogram graph is shown monthly in figure 5.21 and daily in figure 5.22. From the result, it is easily seen that most of the substations have the peak load in July, where the weather is really hot and airconditioner is required, and December, where the weather is cold and where there is the Christmas vacation. The peak load distribution is a good way to classify the substations because different kind of load should has different peak load.

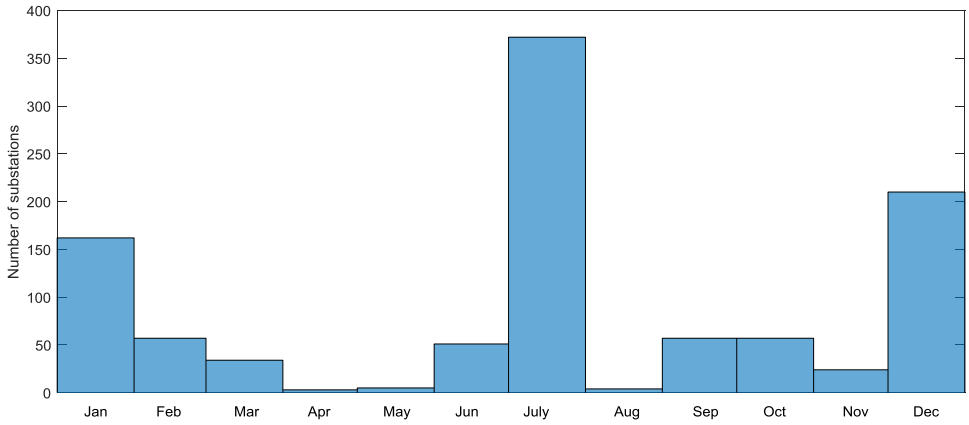


Figure 5.21: The monthly histogram graph for maximum load

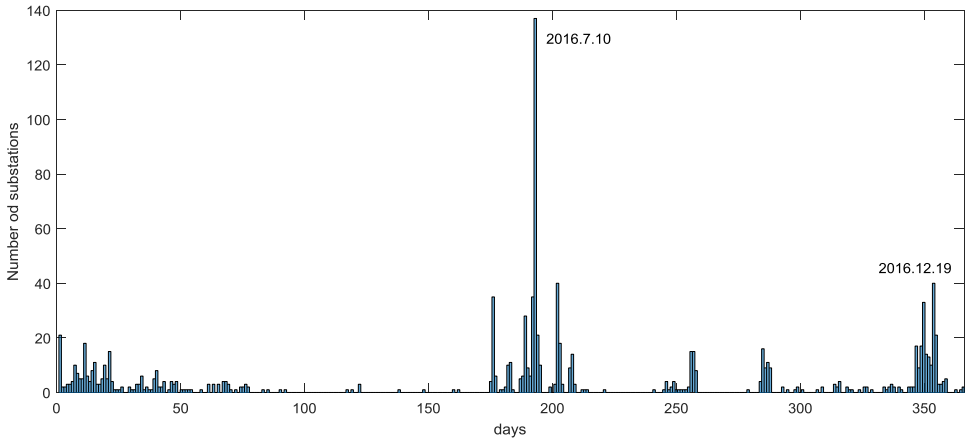


Figure 5.22: The daily histogram graph for maximum load

5.4 Summary

In this chapter, the result of PCA and the result of the load classification are shown after calculation. In PCA method, we extracted 3 components. After analyzing the coefficients of the components and the comparison of load profile with different value of 3 components, the meaning of each component can be supposed: the activities factor, the air conditioner factor and the holiday factor. Then with the result of the PCA, the customers with different load profiles can be classified into 8 types with the octants with new origin in the median. And the result can be shown in the google map, from which we can see there are only 4 types common:

Type 1 customers have the character of uniform activities, high airconditioner occupancy, low holiday differences, for example household out of city center;

Type 2 customers have the character of non-uniform activities, high airconditioner occupancy, low holiday differences, for example the household in city center;

Type 7 customers have the character of non-uniform activities, low airconditioner occupancy, high holiday differences, for example factory or commercial buildings with non-uniform activities;

Type 8 customers have the character of uniform activities, low airconditioner occupancy, high holiday differences, for example factory or commercial buildings with uniform activities.

6 CONCLUSIONS

This thesis mainly discussed about a method to classify distribution load profiles based on PCA method, with the real daily load data from the UNARETI company in Milano. The classification model includes five steps:

a. Loading raw load data from the database. In this step, according to the original database and the goal of the model, the database is refined and simplified.

b. Load data preparation. In this step includes four process. First, defining and deleting the bad data with two indexes: the total missing data and the maximum continuous missing data. Second, normalizing data with the max load normalization after the comparison. Third, the PCHIP+ARMA method is used in the data interpolation. At last, 3-σ rule is applied in outlier detecting.

c. Data simplifying with PCA method. PCA method is used to extract the main components, in order to decrease the dimension of the data.

d. Load profile classification. In this step, we calculated the components and using the octants to classify.

e. Result applications. In this step, the new substation is classified with calculating the main components.

By analyzing the load customers classifying in Milan, some conclusions could be drawn:

1) The first 3 main components of daily load profile are: the activities factor, the air conditioner factor and the holiday factor. Those components contain 68% of the total information in daily load profile in one year. Therefore, to classify the load, these 3 factors are the most important information.

2) There are 8 types of substations after the classification. And after applied to all the Milan substations, it is obvious that the common types are only four: Type 1 customers have the character of uniform activities, high airconditioner occupancy, low holiday differences, and it might be household out of city center; Type 2 customers have the character of non-uniform activities, high airconditioner occupancy, low holiday differences,

and it might be household in city center; Type 7 customers have the character of non-uniform activities, low airconditioner occupancy, high holiday differences, and it might be factory or commercial buildings with non-uniform activities; Type 8 customers have the character of high no-summer power, low airconditioner occupancy, high holiday differences, and it might be factory or commercial buildings with uniform activities.

This model can easily solve the load classification problems. The main innovations in this thesis are:

1) In the data analyzing, the database is totally real, based on the daily data collection in the electricity distribution company in Milan;

2) PCA is applied in the classification to decrease the information, and in the meanwhile the main factors the company should focus on in the daily data are given;

3) In data preparation, the PCHIP+ARMA model is applied creatively, to solve the border problems.

7 REFERENCE

[1] Q. Li, Z. Xu, L. Yang, Recent advancements on the development of microgrids, J. Mod. Power Syst. Clean Energy, vol. 2, no. 3, pp. 206-211, Sep. 2014.

[2].G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, "Customer characterization options for improving the tariff offer", IEEE Trans. Power Syst., vol. 18, no. 1, pp. 381-387, Feb. 2003.

[3] Lin S, Li F, Tian E, et al. Clustering Load Profiles for Demand Response Applications. IEEE Transactions on Smart Grid, 2017.

[4] G. J. Tsekouras, N. D. Hatziargyriou, E. N. Dialynas, Two-stage pattern recognition of load curves for classification of electricity customers, IEEE Trans. Power Syst., vol. 22, no. 3, pp. 1120-1128, Aug. 2007.

[5] N. M. Kohan, M. P. Moghaddam, S. M. Bidaki, G. R. Yousefi, Comparison of modified k-means and hierarchical algorithms in customers load curves clustering for designing suitable tariffs in electricity market, Proc. 43rd Int. Universities Power Engineering Conf., pp. 1-5, Sep. 1—4, 2008.

[6] Z. Zakaria, K. L. Lo, M. H. Sohod, Application of fuzzy clustering to determine electricity consumers' load profiles, Proc. IEEE Int. Power and Energy Conf., pp. 99-103, Nov. 28–29, 2006.

[7] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, P. Jarventausta, Enhanced load profiling for residential customers, IEEE Trans. Power Del., vol. 29, no. 1, pp. 88-96, Feb. 2014.

[8] Tao, Shun, et al. Load forecasting based on short-term correlation clustering. Innovative Smart Grid Technologies-Asia (ISGT-Asia), 2017 IEEE. IEEE, 2017.

[9]G. Chicco, R. Napoli, F. Piglione, M. Scutariu, P. Postolache, C. Toader, Emergent electricity customer classification, Proc. Inst. Elect. Eng. Gen. Transm. Distrib., vol. 152, no. 2, pp. 164-172, Mar. 2005.

[10]G. Chicco, R. Napoli, F. Piglione, Comparisons among clustering techniques for electricity customer classification, IEEE Trans. Power Syst., vol. 21, no. 2, pp. 933-940, May 2006.

[11]S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, F. J. G. Franco, Classification filtering and identification of electrical customer load patterns through the use of self-organizing maps, IEEE Trans. Power Syst., vol. 21, no. 4, pp. 1672-1682, Nov. 2006.

[12]G. Chicco, I. S. Ilie, Support vector clustering of electrical load pattern data, IEEE Trans. Power Syst., vol. 24, no. 3, pp. 1619-1628, Aug. 2009.

[13]J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, M. Mohamad, Nontechnical loss detection for metered customers in power utility using support vector machines, IEEE Trans. Power Del., vol. 25, no. 2, pp. 1162-1171, Apr. 2010.

[14]M Piao, HS Shon, JY Lee, KH Ryu, Subspace Projection Method Based Clustering Analysis in Load Profiling, IEEE Trans. Power Syst., vol. 29, no. 6, pp. 2628-2635, Nov. 2014.

[15] Xiqiao, L., Wanlu, W., Bo, Z., Xu, Y., Shuai, H., & Lijuan, Q. (2018, March). Analysis of large-scale electricity load profile using clustering method. In Networking, Sensing and Control (ICNSC), 2018 IEEE 15th International Conference on (pp. 1-5). IEEE.

[16] Peng, B., Wan, C., Dong, S., Lin, J., Song, Y., Zhang, Y., & Xiong, J. (2016, November). A two-stage pattern recognition method for electric customer classification in smart grid. In Smart Grid Communications (SmartGridComm), 2016 IEEE International Conference on (pp. 758-763). IEEE.

[17] Das, T. K. (2015, October). A customer classification prediction model based on machine learning techniques. In Applied and Theoretical Computing and Communication Technology (iCATccT), 2015 International Conference on (pp. 321-326). IEEE.

[18] Nizar, A. H., Dong, Z. Y., & Wang, Y. (2008). Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems, 23(3), 946-955.

[19] Yang J, Zhao J, Wen F, et al. A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis[J]. IEEE Transactions on Smart Grid, 2018.

[20] Liu B, Qiu H, Shen Y. Establishment and implementation of securities company customer classification model based on clustering analysis and PCA, Control Engineering and Communication Technology (ICCECT), 2012 International Conference on. IEEE, 2012: 325-329.

[21] Ruan, Da; Chen, Guoqing; Kerre, Etienne (2005). Wets, G., ed. Intelligent Data Mining: Techniques and Applications. Studies in Computational Intelligence Vol. 5. Springer. p. 318. ISBN 978-3-540-26256-5.

[22] Gurland J. Hypothesis Testing in Time Series Analysis[J]. 1954

[23] MA Hong-bo, WANG Zheng, LIU Li-ping, et al. Principal Component Analysis and

Calculation of Power Daily Load Data. Journal of Northeast Electric Power, 2017, 38(7): 46-48.

[24] Zhang Peng. Research on Comprehensive Evaluation Based on Principal Component Analysis. Nanjing: Nanjing University of Science and Technology, 2004.