International Journal of Multiphysics
Volume 18, No. 2, 2024
ISSN: 1750-9548

Commercial and Industrial Electricity Consumption Pattern Recognition Using Smart Meter Data: Unsupervised Clustering Method, Application, and Case Study in China

Wei Xiao[1,*], Yuanyuan Hu[2]

[1]State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation College of Management Science, Chengdu University of Technology, Chengdu, Sichuan, 610069, China

[2]Department of Internet Business (Data Center), State Grid Sichuan Electric Power Company Tianfu New District Power Supply Company, Chengdu, Sichuan, 610041, China.

*Corresponding Author.

Abstract

Smart meter deployment in commercial and industrial sector provides an amount of power data that allows analyze the behaviors of commercial and industrial consumers. However, few studies pay attention to commercial and industrial end-users of energy consumption pattern. Besides, conventional methods for load profiling faces the problem of extracting features of smart meter data for making the clustering practical as well as confirming the appropriate clustering parameters. This paper resolves these problems via the unsupervised clustering algorithm that combines significant statistics and behavioral characteristics, and particle swarm optimization and simulated annealing algorithms are implemented to improve the global searching ability of clustering algorithm. A case study of the electricity consumption behaviors of 560 commercial and industrial consumers for a southwest Chinese city is investigated from 2018-2022. Relying on the visual test and validity indices scoring, the algorithm obtains an effective classification of three groups. The results show that electricity consumption can be robustly modeled using improved fuzzy c-means algorithm.

Keywords: Electricity consumption pattern, commercial and industrial consumers, unsupervised clustering, demand side management

1. Introduction

With the emerging of smart grid and advanced metering infrastructure (AMI), various information regarding energy use with high temporal resolution are generated. Data science techniques for treating and analyzing such data shall be deeply studied and implemented, which can effectively assist in improving the insight into consumers' behavior variability, and in continuing the energy revolution in the premise of ensuring reliable electricity supply as well as acceptable grid costs. As Commercial and Industrial Electricity (C&I) consumers accounts for a large proportion of the income of power grid companies, they have higher power supply service requirements than residential customers. However, C&I consumers' electricity consumption behaviors are different with residential consumer [1]. Accurate estimation of commercial and industrial demand is a basic requirement for every C&I customers to find a way to minimize its electricity bill, and will help power supply company optimize the time-of-use (TOU) price scheme and the agent purchasing system. Therefore, there is a need for developing alternative methods for modeling C&I electricity consumption.

Currently, abundant research efforts focus on identifying the patterns for capturing residential electricity structural or behavior dynamics [2-11], and few studies focus on C&I electricity consumption. Various clustering approaches have been adopted clustering the residential electricity consumption. Nevertheless, for

ensuring successful clustering, it is crucially to adopt proper clustering algorithm, correctly select clustering parameters as well as determine the characteristics.

It is possible to divide the work state into two types: (a) the usage of whole data for the clustering and (b) the extraction and usage of certain data characteristics like the peak demand, the load rate. The first type is inapplicable for the smart meter data with high resolution, become the concept of similarity is untenable in the high-dimensional space. The second type plays its role through the quantification of consumers' behavioral and structural characteristics, and the obtained information are used for identifying different consumption profile groups. In the study by Yilmaz, et al. [7], two approaches have been compared regarding the same dataset, finding that characteristics-based clustering exhibits a stronger effectiveness relative to the entire time series. Nevertheless, such extraction of consumption profiles is incapable of fully assessing the variability and the effect on mid-term and long-term demand, and thus the grid operations can not be effectively planned.

Another important aspect of C&I electricity consumption modeling is to develop a model that can handle the mid and long-term behavior features. Most studies only pay attention to the daily load profiles, but the daily electricity consumption data in one month exhibits large difference from the daily load patterns that are based on an hourly or 15-minutes basis. Medium and long time series data present stronger dimensionality, feature correlation and noise, hence the clustering can be difficult to achieve in practice [4].

Based on the arguments presented so far, an improved clustering method is proposed for modeling the C&I electricity consumption to interpret and characterize identified profiles upon the behavioral viability in mid-term and long-term demand period. To better extract features to describe a C&I electricity consumption profile, a preliminary analysis is conducted to identify the consumption data variation at various aggregated levels in electricity consumption in various economic fields, seasons, and months. Such analysis contributes to the identification of the essential behavioral characteristics exhibited by C&I electricity consumers, thereby further serving for the calculation of each consumers' general features. Also, to correctly determine the optimal clustering parameters, simulated annealing (SA) together with particle swarm optimization (PSO) are used for improving the search capability of FCM, namely automatedly searching the optimal fuzzifier, clustering number and the appropriate threshold regarding validity indices. Furthermore, a case study of 560 C&I consumer from a southwest city of China, for a daily electricity consumption data from 2018 to 2022, provided value insights on evaluating the proposed methodology from a long term. The strict data analysis assists in finding the practical features to produce three typical C&I consumption profiling, together with the impact exerted by seasons as well as the potential behavioral characteristics. As found, their consumption behaviors exhibit an obvious heterogeneity in different months and seasons, which made it possible to customize the demand response for better suiting the profiles, and possibly identifying the anomalies.

This paper is organized as follows: section 2 gives the profiling methodology of electricity consumption. Section 3 describes the dataset and conducts empirical analysis for extracting the classical behavior-based clustering features. The profiling results, as well as related analysis and interpretation are provided in section 4. Section 5 is the conclusion of the findings of the study together with some future research.

2. Methods and Models

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

In this section, we first present the conventional FCM algorithm, following with an improved FCM (denoted as IFCM) algorithm integrated with heuristic algorithms. Then, classical clustering validity indices (CVIs) are displayed, and a new validity index is proposed. Finally, a framework of data mining and electricity usage segmentation is developed to explain how we investigate the C&I consumers' electricity consumption pattern.

2.1 Fuzzy-c means clustering (FCM) algorithm

In data clustering, a group of unlabeled data patterns are classified into homogenous clusters considering similarity measure. The fuzzy clustering algorithm produces a fuzzy partition for offering the membership degree

$\mu_{F_i}(x_j)$ of each data point x_j to a certain cluster F_i. The fuzzy clustering method adopts soft-decision pattern in each iteration via the membership functions, hence the possibility for it to generate local minima is smaller than that of the crisp clustering algorithm.

The FCM by Bezdek [12] is the common fuzzing clustering algorithm, classifying a group of pattern data X into c homogenous groups ($\tilde{F}_i, i = 1,2,\dots,c$). Formally, FCM holds the objective of obtaining the fuzzy c-partition $\tilde{F} = \{\tilde{F}_1, \tilde{F}_2, \dots, \tilde{F}_c\}$ for an unlabeled data set $X = \{x_1, x_2, \dots, x_n\}$ and the cluster number c through the minimization of the function J_m,

$$\text{Minimize } J_m(U, V: X) = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m \left\|x_j - v_i\right\| \quad (1)$$

where μ_{ij} stands for the membership degree of data point x_j to the fuzzy cluster $\tilde{F}_i$, and also belongs to a $(c \times n)$ pattern matrix $U = [\mu_{ij}]$. The ith row of U, U_i, represents to a fuzzy cluster $\tilde{F}_i$. $V = (v_1, v_2, \dots, v_c)$ denotes a vector regarding cluster centroids of the fuzzy cluster $(\tilde{F}_1, \tilde{F}_2, \dots, \tilde{F}_c)$. Hence, it is allowed to denote a fuzzy partition by the pair (U, V). $\left\|x_j - v_i\right\|$ represents the Euclidean norm between x_j and v_i. The parameter m takes charge of controlling the membership fuzziness of each datum. It holds the objective of iteratively improving a sequence of sets of fuzzy clusters $\tilde{F}(1), \tilde{F}(2), \dots, \tilde{F}(t)$ until $J_m(U, V: X)$ is not further improved, where t means the iteration step. Algorithm 1 illustrates the FCM algorithm.

Algorithm 1. FCM algorithm

Given a preset cluster number c , a selected value of m, a threshold ϵ, max_iterations r
Initialize $U_i = [\mu_{ij}]$ of x_j belonging to cluster $\tilde{F}_i$ for $i = 1$ to c such that $\sum_{i=1}^{c}\mu_{ij} = 1$.
for $t = 1$ to r **do**

Update the membership matrix $U_i = [\mu_{ij}]$ by using $\mu_{ij} = \left[\sum_{k=1}^{c}\left(\frac{\left\|x_j - v_i\right\|^2}{\left\|x_j - v_k\right\|^2}\right)^{1/(m-1)}\right]^{-1}$.

Calculate the fuzzy cluster centroids V^t by using $v_i = \frac{\sum_{j=1}^{n}(\mu_{ij})^m x_j}{\sum_{j=1}^{n}(\mu_{ij})^m}$.

Calculate J_m^t using Eq.(1)
if $(abs(J_m^t - J_m^{t-1}) < \epsilon$ **then**
break;
else
$J_m^{t-1} = J_m^t$
end if
end

2.2 Optimizing the global searching ability of FCM

FCM clustering acts as a local search algorithm, thus exhibiting limited global search capability. Intelligent algorithms, namely simulated annealing (SA) and particle swarm optimization (PSO) assist in optimizing FCM searching capability. An IFCM algorithm based is developed as illustrated in Algorithm 2

Algorithm 2. IFCM algorithm

Initialize the population *X* in terms of *c, m, T_0, T_c*; // *c* is the number of classifications, *m* is the parameter of FCM, *T_0, T_c* are the threshold parameter.
Initialize *T_{start}, T_{end}, Maxgen, now_ x=X, etc,*. // *T_{start}* is the initialize temperature, *T_{end}* is the final temperature, *Maxgen* is the maximum iterations, *now_ x* is the current particle swarm position.
for *k* = 1 to *Maxgen* **do**
T_k=(T_{end} - T_{start})·(Maxgen - k)/Margen+T_{end}; //*T_k* is the temperature at iteration *k*.
best _score = max (func(best_x)); //Find the optimal value under the current solution, determined by the value of CVIs.
best_x = find(global_score);); // *best_x* is the best position for particles.
If *global_best_score < best_score*:
global_best_score=best_x

local_best_score = best_score
for *m* = 1 to *N: // N* is the size of population of *X*.
for *n* = 1 to *D: // D* is the parameter dimension.
$V_{mn} = w \cdot V_{mn} \cdot rand + c_1 \cdot rand \cdot (best\_x_{mn} - x_{mn}) + c_2 \cdot rand \cdot (global\_best_n - x_{mn}) + randnormal$; //w,$c_1$,$c_2$ are the parameters of the basic PSO algorithm.
$new\_x_m = now\_x_m + V_m$
if $func(new\_x_m) > func(now\_x_m)$ **or** $rand < exp(func(new\_x_m) - func(now\_x_m)/t)$
then: $now\_x_m = new\_x_m$
end
end

IFCM algorithm first initializes the control parameters and population regarding PSO algorithm, and then calculates the fitness value together with the initial membership degree. SA and PSO algorithms are iterated. The algorithm stops with the loop count *gen* reaching the maximal iteration number *Maxgen*, and the annealing temperature *T* below the termination temperature T_{end}. It is capable of reliably and more rapidly searching for the global optima solution as well as restricting the change in position regarding the original and new particles during iteration, meanwhile accelerating the algorithm convergence speed.

Notably, the fuzziness parameter *m* and the weighting exponent *p* are essential for FCM, controlling the cluster number. At present, many research pay attention to fuzzifier value selection. We summarized the key suggestions on the optimal fuzzifier parameter value selection as presented in **Error! Reference source not found.**. Currently, *m* selection in FCM lacks sufficient theoretical guidance or commonly accepted criterion [13-15]. Practically, the selection is completed subjectively by users. Commonly , the used value in practice is *m* = 2 [16-18], which exhibits certain subjectivity. Inappropriately selecting m will remarkably impact the clustering results.

In FCM clustering, the fuzziness parameter *m* is required to be larger than 1 but shall not be too large [16, 18-20]. Thus, the IFCM algorithm sets the initial searching range of *m* to [1.1, 5] considering the existed studies. The FCM clustering algorithm is conducted, where the value of m is different with an increase rate of 0.1. We calculate the average value regarding each CVIs.

2.3 Clustering validity indices (CVIs) for fuzzy clustering

As the clustering acts as an unsupervised learning process, a proper clustering number shall be inevitably determined. The process of finding the most appropriate cluster number is the cluster validation, establishing an indicator function of cluster validity index (CVI), executing the clustering algorithm that has different cluster number in a specific range on the certain data set, and at last identifying the cluster number that corresponds to the optimal CVI value as the optimal cluster number.

The first FCM related validity indices are the partition coefficient (PC) and the partition entropy (PE):

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2 \tag{2}$$

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\log(\mu_{ij}) \tag{3}$$

Weaker fuzziness means larger *PC* value or smaller *PE* value. The *PC* maximization or *PE* minimization regarding the fuzzy cluster number help to obtain the solution partition.

Researchers proposed a lot of validity indices combining the membership degrees and the geometric structure regarding the data set, such as *XB*, *PCAES*. Equation below explains the *XB* and *PCAES(c)* indices:

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m\|x_j - v_i\|^2}{n\left(\min_{i,j=1,\ldots,c,i\neq j}\|v_i - v_j\|^2\right)} \tag{4}$$

$$PCAES(c) = \sum_{i=1}^{c}\sum_{j=1}^{n}\frac{\mu_{ij}^2}{\mu_{Mj}} - \sum_{i=1}^{c} e\left(-\min_{k\neq i}\left(\frac{\|v_i - v_k\|^2}{\beta_T}\right)\right) \tag{5}$$

$$\mu_{Mj} = \min_{1\le i\le c}\sum_{j=1}^{n}\mu_{ij}^2,\ \beta_T = \frac{\sum_{i=1}^{c}\|v_i - \bar{v}\|^2}{c},\ \bar{v} = \sum_{j=1}^{n}\frac{x_j}{n} \tag{6}$$

A larger *PCAES* index indicates strong compactness and good separation for all *c* clusters. Smaller XB index means stronger compactness and better separation.

2.4 A hybrid new CVI considering compactness, overlap and separation measures

At present, various CVIs are put forward specific to different data set, thereinto, no CVI can exhibit a better performance relative to others, and no CVI can be the most applicable for any data sets [21, 22]. Some CVIs only be applicable for certain data type. On that account, designing a proper CVI shall take into account the data set distribution and characteristics. C&I consumers' electricity usage data on a daily basis presents a large cluster size and strong density. Hence, the study proposed a new CVI, namely Cos in this section. For further validating whether the proposed CVI is effective, we implement another four well-known CVIs (e.g., $PE, PC, XB, PCAES(c)$, presented in section 3.3 to evaluate the clustering results.

Generally, indices calculate the variance regarding all data objects in a cluster, for evaluating the compactness. When the variance is bigger, the compactness is lower. $\sum_{j=1}^{n} \mu_{ij}^2 \left\| x_j - v_i \right\|^2$ usually serves for measuring the variance of ith cluster, and the distance between data objects and cluster centers as well as the membership value that describes fuzziness are considered. These measure monotonic falls to 0 as the cluster number c increases, $\lim_{c \to n} \left\| x_j - v_i \right\| = 0$. Hence, it is incapable of validating the partitions that have many small clusters. Traditional compactness measurement also exhibits another shortcoming, i.e., it wrongly estimates the compactness regarding two clusters which have the same element numbers and distributions while the sample sizes are different [23].

There are also challenges facing separation measures. The most common cluster separation measure can be obtained through the calculation of the distance between the centers of clusters $\left\| v_i - v_j \right\|$. Nevertheless, the separation measure fails to obtain the accurate cluster separation value [23].

For overcoming these challenges, a new index is given, estimating three properties regarding fuzzy clusters: the compactness, the overlap, and the separation. The compactness serves for measuring the data variation or scattering. Smaller scattering is accompanied by higher compactness. The overlap serves for measuring the overlap degree between fuzzy clusters. Smaller overlap indicates larger separation and the clear assignment of each data object to only one cluster. The proposed validity index has a basic goal of finding partitions with the maximized compactness and minimized overlapping. First, we present the definition of the compactness measure and overlap measure.

Definition 1. The fuzzy partition compactness degree $C(c, U)$ refers to compactness degrees regarding all cluster $C_i(c, U)$. The compactness degree regarding each cluster is a sum of compactness rates regarding all data objects $C_{ij}(c, U)$*:*

$$\begin{aligned} &C(c,U) = \sum_{i=1}^{c} C_i(c,U),\ C_i(c,U) = \frac{1}{n}\sum_{j=1}^{n} C_{ij}(c,U) \\ &C_{ij}(c,U) = \begin{cases} \mu_{ij} & \text{if } (\mu_{ij} - \mu_{ik}) \geq T_c,\ \ k = 1, \dots, c, \quad k \neq l \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

The *j*th data point exhibits increasing compactness regarding the *i*th cluster in the *i*th cluster, that is μ_{ij} is larger than T_c in terms of all other membership values that describe the association degree of this data objects to other clusters. T_c is the threshold.

Definition 2. The overlap measure $O_{ab}(c, U: C_a, C_b)$ between two clusters C_a and C_b is computed from overlap degrees $O_{abj}(c, U: C_a, C_b)$ of each data object x_j, that exhibits a strong enough association with both fuzzy clusters C_a and C_b. A small value of overlap measure $O_{ab}(c, U: C_a, C_b)$ between two fuzzy clusters C_a and C_b indicates their small overlap and good separation. The threshold T_o serves for eliminating the noise points on the boundary of cluster, representing the separateness between samples of two clusters C_a and C_b.

$$\begin{aligned} &O_{ab}(c,U) = \frac{1}{n}\sum_{j=1}^{n} O_{abj}(c,U), \quad a,b = 1, \dots, c, \quad a \neq b \\ &O_{abj}(c,U) = \begin{cases} 1 - abs(\mu_{aj} - \mu_{bj}) & \text{if } abs\,(\mu_{aj} - \mu_{bj}) \geq T_o \text{ and } a \neq b \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

The proposed summation type index *Cos* is defined as the calculation of compactness and separation with appropriate weighting factor.

$$\begin{aligned} Cos(c,U) &= [C(c,U) - O(c,U)] \times S(c,U) \\ &= \left[\frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{c} C_{ij}(c,U) - \sum_{a=1}^{c-1}\sum_{b=a+1}^{c} O_{abj}(c,U)\right)\right] \times \min_{1\le i,j\le c, i\ne j}\left\|v_i - v_j\right\| \end{aligned} \tag{9}$$

From Eq.(9), we can see that a good clustering partition should ensure large enough compactness and separation between various groups, and small enough overlapping among different group, i.e., bigger *Cos* index value is accompanied by proper clustering result. On that account, with the cluster number changing, the maximum *Cos* index value decides the most proper cluster number.

2.5 Model framework

To mine the C&I consumers' electricity usage profiles, Figure 1 gives a mining structure, which is composed of three modules, data acquisition, data processing and analysis, electricity usage pattern recognition.

The data acquisition module collects as well as stores the electricity consumption data. And the second module processes raw time-series data for handing the missing values, as well as filters the yearly data for currently capturing the impact brought about by seasons. Then an exploratory analysis is carried out regarding the consumption data, for visualizing the increasing consumption amount in months, seasons and years. It analyzes the dataset in detail for identifying the major features influencing end-users' consumption behavior. It exhibits a large significance, ensuring the correctness of data and features that are adopted in the subsequent modules.

The third module is to mine the electricity usage behaviors and obtain the potential regulations. This part essentially supplements the whole process. The module discusses fuzziness parameters and the optimal number of clustering based on the IFCM algorithm, and then identifies as well as characterizes the typical and atypical consumption profiles.

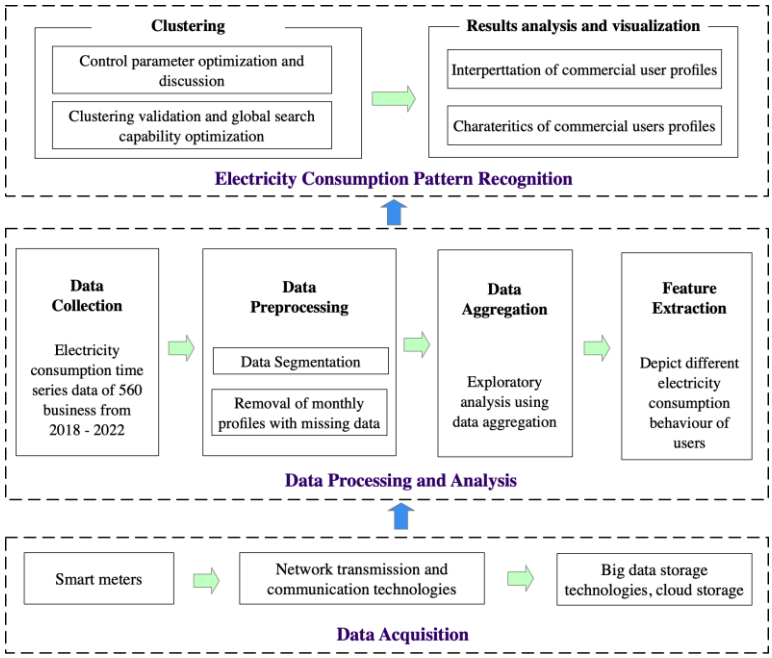


Figure 1 Framework of the proposed model.

3. Data

The dataset contains electricity consumption data from the 560 C&I consumers in Tianfu New District, located in Chengdu, Sichuan of southwest China, with a 1-month resolution. The data collection lasts from 2017 to 2022. This data is private, and came from the State Grid Tianfu New Area Electric Power Supply Company which has processed and pseudonymized the data before research.

Chengdu Tianfu New District plays a big role for the development a regional center and a greater metropolis. It is a major component for opening the Chinese inland areas to Europe and other parts of Asia, and can assist in pushing the city's modern industry development as well as fueling the entire west's economic development. These parts cover three cities, seven counties and 37 townships. The underlying concept proves its status in China and brings into a crucial investment opportunity. The geographical location of Chengdu Tianfu New area of Sichuan Province in China is shown in **Error! Reference source not found.**.

We first removed the profiles which have missing values and maximum zeros, and considered the yearly demand for corresponding to seasonal influence on the consumption, and then obtained a smaller data set covering the period from March 2018 to May 2022.

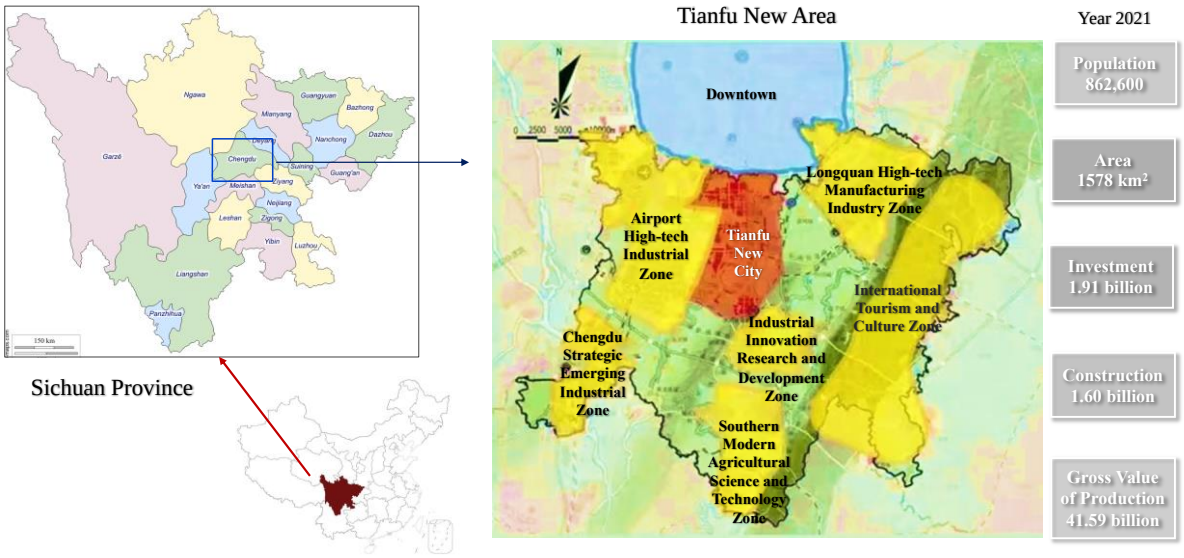
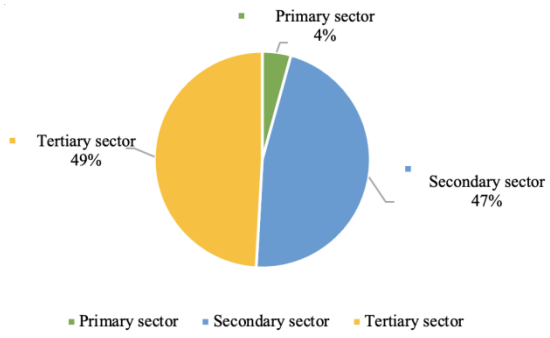


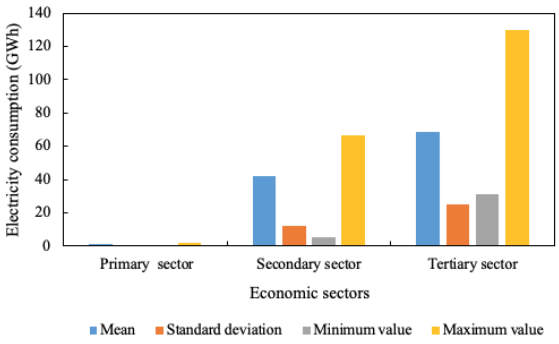
Figure 2 Location of China Tianfu New Area.

4. Results and Discussion

4.1 Empirical analysis



(a) Industrial structure



(b) Descriptive statistics of the electricity consumption by sectors

Figure 3 The Status quo of Tianfu New Area

Empirical analysis will be first performed on data set for visualizing the time series regarding electricity consumption data. Hence, the dataset has been processed as well as aggregated for the data analysis in different periods: month, season, and year.

According to the statistics of annual economic report of Tianfu New area, the industrial structure of Tianfu New Area is presented in **Error! Reference source not found.**a. It clearly shows that the share of the secondary and tertiary industry comes in at 96% where tertiary sector attributes nearly half of the Gross Domestic Product

(GDP). This analysis identifies main industries of the C&I electricity consumption consumers. **Error! Reference source not found.**b displays the basic statistics of the three sectors over 5 years where tertiary sectors saw the highest consumption, and the primary sector reported the smallest consumption.

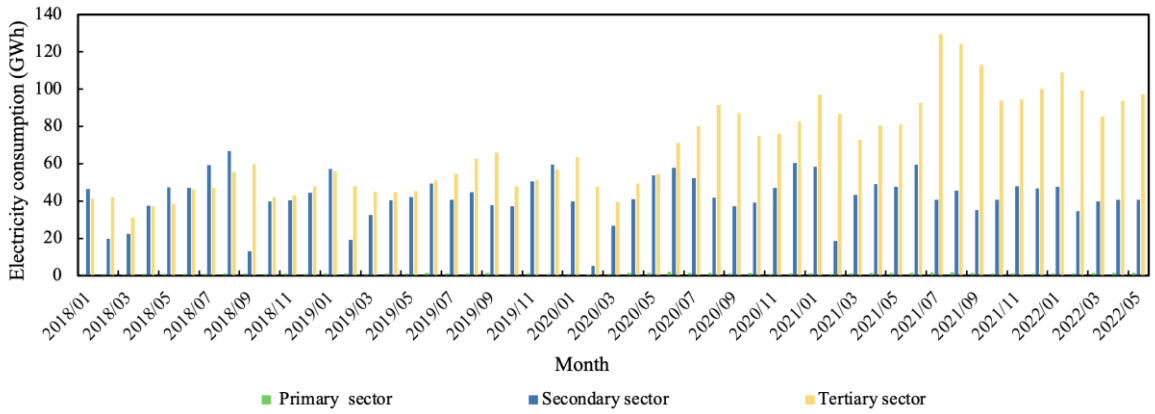
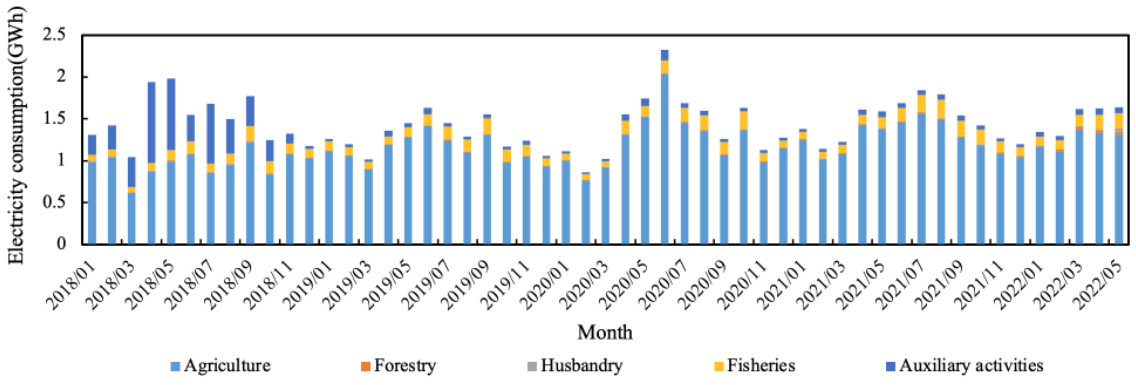


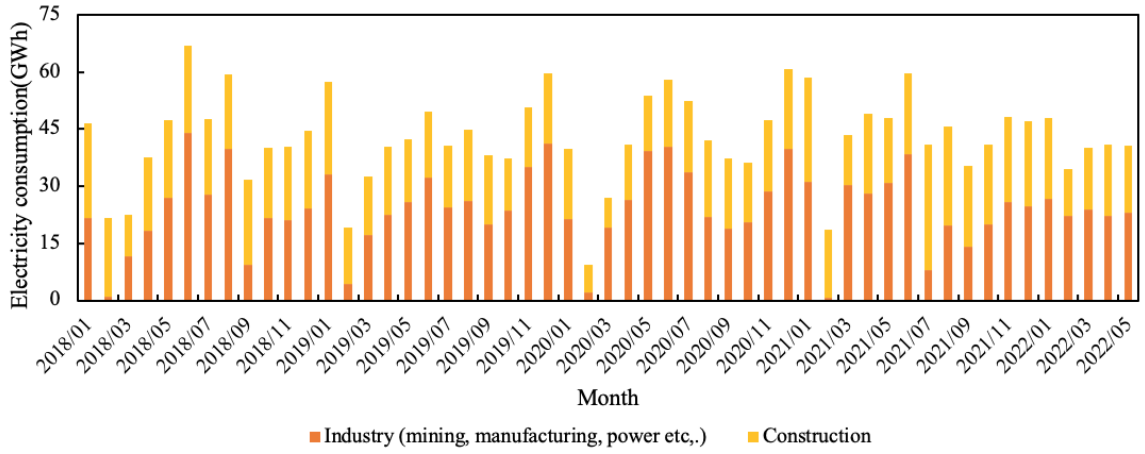
Figure 4 Proportion of monthly electricity consumption of different economic sectors.

Figure 4 depicts the monthly electricity consumption of different sectors on demand profiles. As the trend shows the primary and secondary sectors almost keep the stable consumption level from 2018 to 2022, whereas the consumption profile of tertiary sector has an obvious climbing trend during years, a highest demand in the third quarter of 2021. As shown by the graph, the secondary sector presented the highest demand from November to January every year, whereas that of the tertiary sector was seen from July to September. Hence, demand in the summer and winter seasons was the highest, then the autumn and spring.

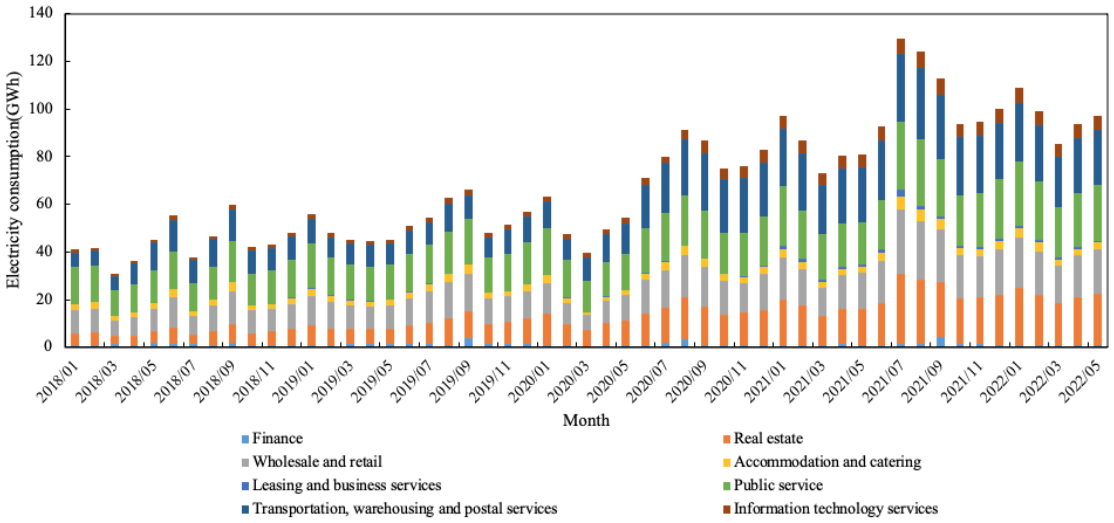
For better visualizing the demand profile, the sectors are divided into various industries. Figure 5 depicts the industrial electricity consumption of each sector in details. Figure 1a observes the consumption of agriculture comes in at 60% of the total consumption of the primary sector where the highest demand was observed from April to October, whereas the lowest demand could be seen from January to March. Figure 1b observes consumption profile of mining and construction industries with a slightly stronger granularity for the later spring and summer and winter late autumn and winter. Figure 1c clearly shows that the finance, public service, and information service-related industries have a higher demand from June to September and January to March. Figure 1 confirms that demand in summer is the highest, and then winter, spring, and autumn.



(a) Primary sector



(b) Secondary sector



(c) Tertiary sector by specific industries

Figure 1 Monthly electricity consumption.

Above graphs illustrates different consumption intensities in different sectors and industries. Results indicate that Tianfu New Area is characterized by the secondary and tertiary features, and thus this study investigates the electricity usage pattern of C&I and industrial consumers. Industries' consumption profiles obey certain pattern and exhibit a seasonality, possible affected by the weather condition, business, or working or non-working days. The paper focuses on the analysis from the perspective of clustering, combining C&I customers who have similar behavior profiles into groups. We hope this study provides some helpful insights that robust across further load prediction and pricing experiments via the characterization and categorization of the electricity demand response following conventional business classification.

The features for consumption profile clustering will be detailed later, aiming at accurately capturing the consumption patterns in different months as well as confirming the monthly and yearly usage variability.

4.2 Feature extraction and selection

For an effective capturing of C&I customers' heterogenous consumption behavior, data mining is conducted regarding the dataset, for identifying and characterizing seven attributes (

Table 1 Feature used for clustering.

Features	Definition	Description

Annual average consumption	$\left(\frac{1}{12}\right)\sum_{t=1}^{12}$ montly electricity consunption	Annual average monthly electricity consumption
Annually standard deviation	$\left(\frac{1}{12}\right)\sum_{t=1}^{12}\text{abs}\begin{pmatrix}\text{electricity consumption at a month}\\-\text{annual average consumption}\end{pmatrix}$	Standard deviation of annual monthly electricity consumption.
Monthly consumption CV	$\frac{\text{Standard deviation of montly electricity consunption}}{\text{Mean of montly load}}$	Coefficient of variation in monthly consumption capturing the variation of between-months monthly consumption.
Monthly consumption SCV	$\sum_{t=1}^{12}\frac{\text{Standard deviation of montly electricity consunption}}{\text{Mean of montly load}}$	Sum of coefficients of variation in annually consumptions capturing the total magnitude of yearly variation of load in a year.
Monthly mean electricity consumption	$\left(\frac{1}{N}\right)\sum_{i=1}^{N}$(montly electricity consunption)	Average monthly electricity consumption of all data.
Monthly mean standard deviation	$\left(\frac{1}{N}\right)\sum_{i=1}^{N}\begin{matrix}\text{abs}(m\text{onthly electricity consumption}\\-\ m\text{onthly mean electricity consumption})\end{matrix}$	Standard deviation of monthly electricity consumption of all data.
Area	The build-up area of each building	Only the floor area is recorded, in square kilometers
Business type	Classified according to the ISIC code	Including Healthcare, Warehousing, Public service, etc.

). The calculation of these attributes is on a yearly basis, taking into account the period, and the measurement about month-to-month variability and fluctuation regarding every type of C&I building. They can serve for different datasets which have various temporal resolutions, and can be easily computed. Notably, the entire consumption data are carefully normalized before calculating these attributes.

Besides, another combination of attributes is also considered, like the seasonality for every time period, as well as the monthly entropy specific to seasons and years. Nevertheless, there is a strong correlation between all these attributes and the monthly mean electricity consumption and the seasonality. Hence, for reducing the attribute dimensionality and correlation, we only select those that have been listed.

Table 1 Feature used for clustering.

Features	Definition	Description
Annual average consumption	$\left(\frac{1}{12}\right)\sum_{t=1}^{12}$ montly electricity consunption	Annual average monthly electricity consumption
Annually standard deviation	$\left(\frac{1}{12}\right)\sum_{t=1}^{12}\text{abs}\begin{pmatrix}\text{electricity consumption at a month}\\-\text{annual average consumption}\end{pmatrix}$	Standard deviation of annual monthly electricity consumption.
Monthly consumption CV	$\frac{\text{Standard deviation of montly electricity consunption}}{\text{Mean of montly load}}$	Coefficient of variation in monthly consumption capturing the variation of between-months monthly consumption.
Monthly consumption SCV	$\sum_{t=1}^{12}\frac{\text{Standard deviation of montly electricity consunption}}{\text{Mean of montly load}}$	Sum of coefficients of variation in annually consumptions capturing the total magnitude of yearly variation of load in a year.
Monthly mean	$\left(\frac{1}{N}\right)\sum_{i=1}^{N}$(montly electricity consunption)	Average monthly

electricity consumption		electricity consumption of all data.
Monthly mean standard deviation	$\left(\frac{1}{N}\right)\sum_{i=1}^{N}$ abs($m$onthly electricity consumption $-$ $m$onthly mean electricity consumption)	Standard deviation of monthly electricity consumption of all data.
Area	The build-up area of each building	Only the floor area is recorded, in square kilometers
Business type	Classified according to the ISIC code	Including Healthcare, Warehousing, Public service, etc.

4.3 Clustering results

The IFCM (Algorithm 2) partitioned data for obtaining the maximum value of validity index *Cos*, assisting in finding the solution fuzzy c-partitioning and the cluster number c^*. For IFCM, we used the *eps*=0.01, iteration =100, m, T_o, and T_c are unknown parameters and will be solved with IFCM as well. While IFCM exhibits a sensitivity to initialization values, IFCM cluttering is repeated for ten times for each cluster number *c* and the validity index *Cos* is calculated.

We applied the IFCM algorithm and the appropriate parameters are presented in**Error! Not a valid bookmark self-reference.**. After 30 iterations, the new validity index *Cos* sees its maximum (15.35) when c^*= 3 and the fuzziness m =1.01, T_0 and T_c equal to 0.68 and 0.73, respectively. To further verify the effectiveness of proposed validity index, four well-known CVIs (*XB, PECAS, PE, PC*) serve for grouping result evaluation.

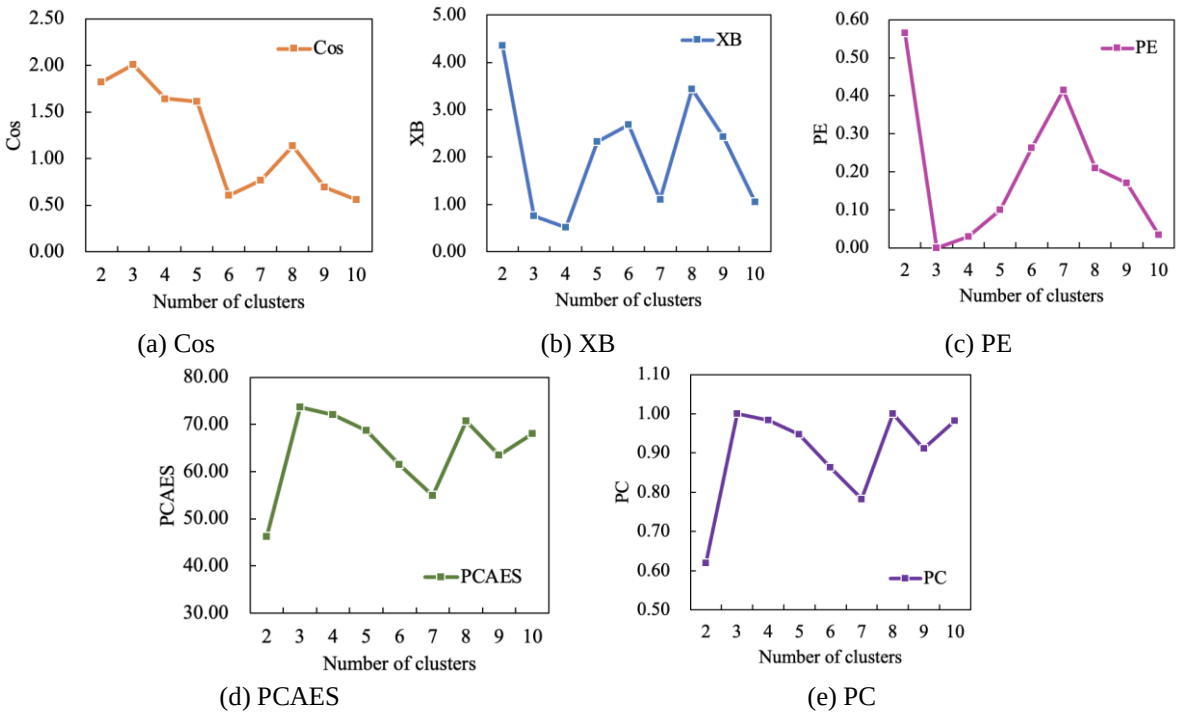
Table 3 lists the values for *XB, PECAS, PE* and *PC* for the electricity consumption data. When the clusters number changes from 2 to 10, c = 3 and has the maximum points of *Cos*, *PECAS*, *PC* curves and the minimum point of *PE* (Figure 2), whereas in view of *XB*, c = 3 is the second reasonably cluster number. Generally, based on the results of a majority of validity indices, the optimal cluster number c^*= 3. Additionally, the proposed *Cos* index shows that c^*= 3 is a proper clustering number for estimating the monthly electricity usage data, which proves its effectiveness as those classical CVIs.

Table 2 Optimal parameters by FCM algorithm with validity index Cos.

Parameters	c^*	m	T_o	T_c
Optimal value	3	1.01	0.68	0.73

Table 3 CVIs values for optimal clustering parameters for electricity consumption.

Parameters	PCAES	XB	PC	PE
c^*	3	4	3	3
m	1.05	2.7	1.01	1.03
T_o	0.72	0.5	0.66	0.62
T_c	0.41	0.3	0.25	0.31



(a) Cos (b) XB (c) PE

(d) PCAES (e) PC

Figure 2 Values of CVIs with different number of clusters.

Notably, it is very complex to confirm a proper cluster number for unsupervised clustering. In real-world cases, determining factors of the cluster number are quantitative calculation together with the actual needs for the problems. Specific to a given dataset with n data objects, the cluster number is required to be [2, $n^{-1/2}$][21].

Table 4 lists briefly statistical indicators of groups. Clearly, group 1 exhibits the maximum average monthly electricity consumption, while group 3 exhibits the lowest. In view of the standard deviation, group 1 and group 2 have larger values further demonstrated that group 1 has the highest volatility, indicating the larger sensitivity of these two groups to the external environment. Hence, it seems that demand response programs based on price and incentive the more suit for these kinds of businesses.

Table 4. Statistical indicators of each group (c^*= 3) in GW.

Statistics	Group 1	Group 2	Group 3
Mean	9650.74	2525.36	672.39
Standard deviation	10728.69	1790.48	746.34
Minimum value	959.52	325.53	6.14
Maximum value	23438.16	5892.42	2719.46

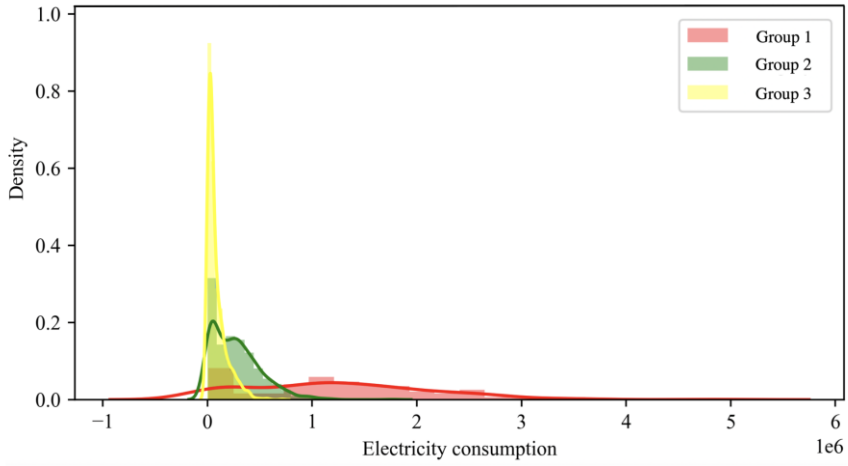
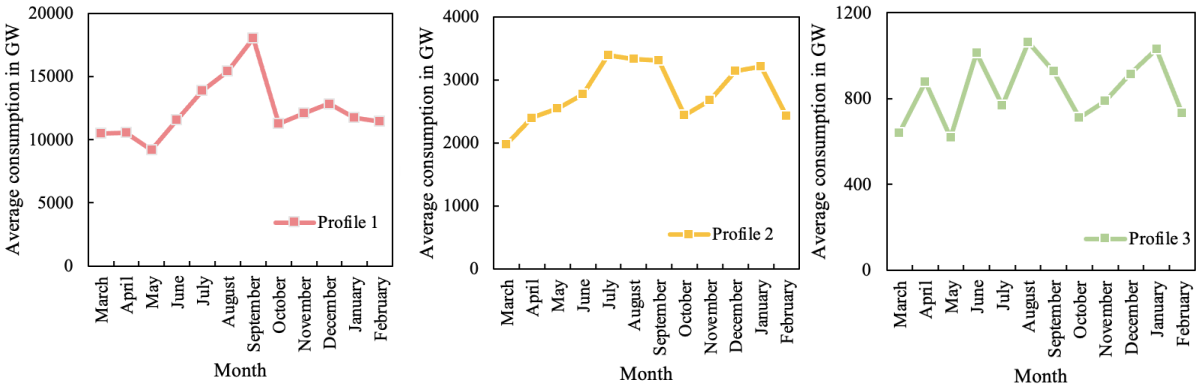


Figure 3 Visual analysis of kernel density estimation (KDE) distribution analysis of groups.



(a) Profile 1 (b) Profile 2 (c) Profile 3

Figure 4 Characterization of primary consumption profiles.

The kernel density estimation (KDE) of each clustering and corresponding typical consumption profiles are plotted in Figure 3 and Figure 4. KDE function is implemented to capture the shape of cross-sectional distributions of electricity consumption intensity of groups. These shapes, based on the feature calculation, represent the consumption profile regarding C&I customers based on increasing consumption information in related time period, revealing the consumption variability and features which contribute to the demand management.

From Figure 3 and Figure 4, we could observe three primary C&I consumers' shapes using clustering. The end-use consumption profile stands for the monthly electricity usage routine usually employed by the C&I customers. Three shapes of representative consumption as observed below:

- Profile 1 (group 1): This first profile, group 1 shows present highest dispersion in overall levels of consumption (Figure 3). This profile features the lowest consumption during the spring, which continuously increases with the summer and reaches its peak (around 17000GW) at late summer early autumn (Figure 4). The spring and winter present relatively stable consumption relative to other seasons considering the average profile (9000 – 13500 GW).

- Profile 2 (group 2): The second profile, group 2, presents modest dispersion in overall levels of consumption (Figure 3). This profile features a significant low consumption in early spring late winter, but the consumption keeps increasing from spring, reaches the peak consumption (around 3300 GW) in summer and remains high during this season, then a sharp decreasing in the early autumn and display an obviously increasing trend from the mid-autumn as shown in Figure 4.

- Profile 3 (group 3): This profile displays the lowest dispersion of electricity consumption compared the above two profiles. However, the profile shows a larger variability in the behavior over the year, that reports multiple peak consumptions and month-to-month behavior change whereas all peaks are fluctuated around 1000GW. Hence it is difficult to confirm the peak demand period for the profile. Hence, the variability is caused by highly changing consumption intensity during various seasons. The profile is a combination of consumers whose consumption levels are obviously high and obviously low. Peak consumption in summer and winter are observed as well in group 3, along with a relatively low demand in spring.

Figure 5 displays the share of different C&I consumers of the summarized profiles within business type defined by ISIC code. High-variability clusters i.e., profile 3, collectively takes up nearly 65% of the C&I consumers and 35% of the remaining consumers exhibit low variability i.e., profiles 1 and 2. The surprising results is that the heterogeneity is pronounce even for the same 2-digit ISIC business sectors, hence there is no single variability clusters that focus on characterizing the sectors from the statistically significant level. The consumer shares regarding the individual clusters also change in different sectors, and many exhibit the full spectrum of variability, with the pattern from being randomly fluctuated to flat stable. Even for food, waste industries (Public Administration) with variable operations, not all of these businesses show random fluctuation patter, about 3% remain in the flat stable pattern.

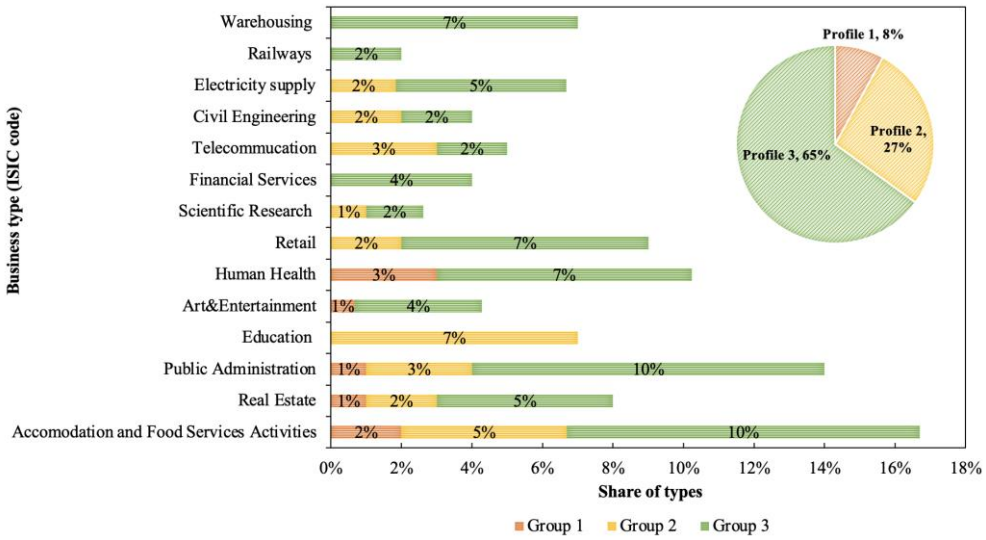


Figure 5 Proportion of Business types (ISIC code) of different profiles.

Overall, profile 1 differs from both due to a solely significant peak consumption whereas the other two reports its peak shifting from summer to winter. The profiles 1 and 2 display obvious seasonality and the behavior is relatively regular considering the peak demand while profile 3 reports a frequent peak shift from seasons. This indicates more than 50% C&I business display less seasonality, but the reasons result in such random fluctuation pattern need investigated in the future work.

5. Conclusion

This paper develops a novel algorithm for analyzing the behavioral characteristics regarding C&I end-users, thereby identifying and characterizing their mid- and long-term heterogenous electricity consumption behaviors. Method evaluation is conducted based on the electricity consumption data from 560 C&I consumers of a city in southwestern China. A generally preliminary analysis is presented to describe the industry structure and the status quo of electricity consumption level by sectors (primary, secondary, and tertiary). Different types of electricity consumption features including various statistics (e.g., the mean consumption, mean standard deviation, coefficient of variation, etc.), area and business type, are considered for C&I end-user side. Different validity indices are implemented to verify the optimal clustering number and the effectiveness of the new validity index. Typical C&I electricity consumption behaviors are identified upon the mid-and long-term characteristics, corresponding patterns are summarized as well.

The approach assists in remarkably weakening the time series dimensionality through the extraction of suitable behavior features, thereby offering concisely represented electricity usage profiles based on the long time series of consumption data. Relying on such simplified data, our approach is more applicable for processing large-scale data. On the other hand, the methodology can serve for improving the peak load prediction specific to a power system zone. The prediction of the total peak load in a certain month or season can only be affected by a subset of the commerce set that belongs to a related class. Hence, the prediction accuracy can be obviously increased by collecting additional information of such commerce.

The future improvement may include the pattern recognition by high-frequency electricity load data with a resolution of 15 minutes. Another improvement relates to consider the impacts from exogenous variables such as weather, COVID-19 pandemic.

Data Statement

The data that support the findings of this study are available under request. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from Bing He with the permission of State Grid Sichuan Electric Power Company Tianfu New District Power Supply Company.

Acknowledgment

This paper is supported by State Grid Sichuan Electric Power Company Tianfu New District Power Supply Company under [grant number SGSCTF00YYJS2100110].

References

[1] M.U.Nwachukwu, N.F.Ezedinma, and U.Jiburum, "Comparative Analysis of Electricity Consumption among Residential, Commercial and Industrial Sectors of the Nigeria's Economy," Journal of Energy Technologies and Policy vol. 4, no. 3, pp. 2224-3232, 2014.

[2] R. Kaur and D. Gabrijelčič, "Behavior segmentation of electricity consumption patterns: A cluster analytical approach," Knowledge-Based Systems, vol. 251, p. 109236, 2022/09/05/ 2022, doi: https://doi.org/10.1016/j.knosys.2022.109236.

[3] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," Applied Energy, vol. 135, pp. 461-471, 2014/12/15/ 2014, doi: https://doi.org/10.1016/j.apenergy.2014.08.111.

[4] S. Aghabozorgi and T. Y. Wah, "Clustering of large time series datasets," Intelligent Data Analysis, vol. 18, pp. 793-817, 2014, doi: 10.3233/IDA-140669.

[5] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series," Applied Energy, vol. 237, pp. 11-24, 2019/03/01/ 2019, doi: https://doi.org/10.1016/j.apenergy.2018.12.063.

[6] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," Renewable and Sustainable Energy Reviews, vol. 120, p. 109628, 2020/03/01/ 2020, doi: https://doi.org/10.1016/j.rser.2019.109628.

[7] S. Yilmaz, J. Chambers, and M. K. Patel, "Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management," Energy, vol. 180, pp. 665-677, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.energy.2019.05.124.

[8] G. Trotta, "An empirical analysis of domestic electricity load profiles: Who consumes how much and when?," Applied Energy, vol. 275, p. 115399, 2020/10/01/ 2020, doi: https://doi.org/10.1016/j.apenergy.2020.115399.

[9] E. L. Ofetotse, E. A. Essah, and R. Yao, "Evaluating the determinants of household electricity consumption using cluster analysis," Journal of Building Engineering, vol. 43, p. 102487, 2021/11/01/ 2021, doi: https://doi.org/10.1016/j.jobe.2021.102487.

[10] T. Yang, M. Ren, and K. Zhou, "Identifying household electricity consumption patterns: A case study of Kunshan, China," Renewable and Sustainable Energy Reviews, vol. 91, pp. 861-868, 2018/08/01/ 2018, doi: https://doi.org/10.1016/j.rser.2018.04.037.

[11] M. Amayri, C. S. Silva, H. Pombeiro, and S. Ploix, "Flexibility characterization of residential electricity consumption: A machine learning approach," Sustainable Energy, Grids and Networks, vol. 32, p. 100801, 2022/12/01/ 2022, doi: https://doi.org/10.1016/j.segan.2022.100801.

[12] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition). Springer US, 2013.

[13] Y. Jian, C. Qiansheng, and H. Houkuan, "Analysis of the weighting exponent in the FCM," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 1, pp. 634-639, 2004, doi: 10.1109/TSMCB.2003.810951.

[14] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," IEEE Transactions on Fuzzy Systems, vol. 3, no. 3, pp. 370-379, 1995, doi: 10.1109/91.413225.

[15] K. Zhou, S. Yang, and Z. Shao, "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study," Journal of Cleaner Production, vol. 141, pp. 900-908, 2017/01/10/ 2017, doi: https://doi.org/10.1016/j.jclepro.2016.09.165.

[16] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, no. 2, pp. 191-203, 1984/01/01/ 1984, doi: https://doi.org/10.1016/0098-3004(84)90020-7.

[17] L. O. Hall, A. M. Bensaid, L. P. Clarke, R. P. Velthuizen, M. S. Silbiger, and J. C. Bezdek, "A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain," IEEE Transactions on Neural Networks, vol. 3, no. 5, pp. 672-682, 1992, doi: 10.1109/72.159057.

[18] S. Yi, S. Hong, and Z. Jian Qiu, "Improvement and optimization of a fuzzy C-means clustering algorithm," in IMTC 2001. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No.01CH 37188), 21-23 May 2001 2001, vol. 3, pp. 1430-1433 vol.3, doi: 10.1109/IMTC.2001.929440.

[19] I. Ozkan and I. B. Turksen, "Upper and lower values for the level of fuzziness in FCM," Information Sciences, vol. 177, no. 23, pp. 5143-5152, 2007/12/01/ 2007, doi: https://doi.org/10.1016/j.ins.2007.06.028.

[20] K.-L. Wu, "Analysis of parameter selections for fuzzy c-means," Pattern Recognition, vol. 45, no. 1, pp. 407-415, 2012/01/01/ 2012, doi: https://doi.org/10.1016/j.patcog.2011.07.012.

[21] M. Kim and R. S. Ramakrishna, "New indices for cluster validity assessment," Pattern Recognition Letters, vol. 26, no. 15, pp. 2353-2363, 2005/11/01/ 2005, doi: https://doi.org/10.1016/j.patrec.2005.04.007.

[22] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2095-2117, 2007/10/01/ 2007, doi: https://doi.org/10.1016/j.fss.2007.03.004.

[23] K. Rizman Žalik, "Cluster validity index for estimation of fuzzy clusters of different sizes and densities," Pattern Recognition, vol. 43, no. 10, pp. 3374-3390, 2010/10/01/ 2010, doi: https://doi.org/10.1016/j.patcog.2010.04.025.