Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China

DOI:

[10.1016/j.energy.2018.09.156](https://doi.org/10.1016/j.energy.2018.09.156)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Fu, X., Zeng, X., Feng, P., & Cai, X. (2018). Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. *Energy*, 165(B), 76-89.
<https://doi.org/10.1016/j.energy.2018.09.156>

Published in:

Energy

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China

Xin Fu^{a,*}, Xiao-Jun Zeng^b, Pengpeng Feng^a, Xiuwen Cai^c

^a*Department of Management Science, School of Management, Xiamen University, Xiamen, 361005, China*

^b*Department of Computer Science, University of Manchester, M13 9PL, U.K*

^c*State Grid Quanzhou Electric Power Supply Company, Quanzhou, 362000, China*

Abstract

The introduction of a new pricing mechanism, the increasing-block tariff (IBT), will not only affect electricity bills for residents, but also lead to a change in residential electricity consumption behaviours. Understanding these consumption patterns will help create more accurate load forecasting and increase the efficiency of the IBT. This study proposes an innovative clustering-based approach for short-term load forecasting under the IBT in China. The new approach initially partitions households into homogeneous groups each of which has distinctive consumption patterns under the IBT, each consumer segment can then select the most appropriate model for load forecasting, and the predicted load demands of different clusters are aggregated to derive the total usage. In particular, the IBT-related attributes are newly introduced into the clustering analysis. The utility and effectiveness of the proposed model is confirmed through a realistic dataset that contains the daily household-level consumption data of 533 households from April 2014 to February 2015. Consequently, the households are classified into five clusters with distinctive consumption patterns, including low-demand and insensitivity to high temperature ($Cluster_1$), ordinary users and sensitivity to high temperature ($Cluster_2$), ordinary users and sensitivity to the IBT ($Cluster_3$), high-demand consumers and sensitivity to high temperature ($Cluster_4$), and luxury consumers ($Cluster_5$). In addition, the obtained experimental results demonstrate that the proposed approach can not only achieve better prediction accuracy (e.g., the mean absolute

percentage error (MAPE) improves from 3.82% to 2.28% by using autoregressive integrated moving average (ARIMA)), but also provide better flexibility for hybrid modelling. From the practical implication point of view, the proposed forecasting model can help power companies to provide a reliable and high-quality electricity supply as well as to establish appropriate schedules of operations and maintenance within a certain area. Moreover, the identified consumption behaviours can be analysed and used to improve the design and promote awareness/acceptance of the IBT.

Keywords:

Increasing-block tariff (IBT), Clustering analysis, Short-term load forecasting, Smart meters

*Corresponding author, Email: xfu@xmu.edu.cn

1. Introduction

With the rapid development of the Chinese economy and the rising income of residents, electricity consumption has increased dramatically in recent years, bringing challenges for electricity supply and storage and raising environmental concerns. For a long time, China has implemented a low-price flat tariff in residential sectors to relieve residents' living burdens; however, this tariff has been heavily subsidised by the industrial and commercial sectors. The drawbacks of this low residential tariff have gradually become evident, the inefficiency and unfairness caused by this cross-subsidies has been greatly discussed in [36, 38, 52, 55]. It is demonstrated in [35] that the 22% poorest residents receive only 10.1% of the electricity subsidy, whereas the top 27% of high-income residents enjoy 45% of the subsidy. These nontargeted universal subsidies run counter to the initial subsidy objective, because they not only lead to inefficiency in cross-subsidy allocation and undermine social equity and stability, but also hinder electricity marketisation reform.

To address rising energy demands and improve the current tariff structure, the Chinese government has attached great importance to promoting the reform of the residential electricity tariff. A new pricing mechanism for electricity, the increasing-block tariff (IBT), has been advocated and implemented in China. IBT has been widely used around the world (e.g., in the United States [50] and Japan [47]) and is a form of price discrimination [37, 48]. In principle, IBT charges higher electricity prices for higher electricity consumption. IBT designs several blocks for household electricity consumption; each consumption block is assigned a certain price, and the price increases with electricity consumption. When electricity consumption exceeds the upper bound of the current block, the higher price of the next block will be charged instead. A reasonable IBT has positive impacts on residential welfare, social equality and efficiency, allocation of cross-subsidies, and energy conservation [37, 52]. More specifically, IBT can relieve financial pressure on low-income residents by charging a lower price for the first block, which is usually designed to ensure essential living needs, while

encouraging energy saving by setting a higher price for luxury electricity consumption [55]. In other words, this mechanism can help to generate more efficient allocation of electricity resources and to transfer the cross-subsidies from high-income residents to low-income households. Having recognised the features and benefits of IBT, this mechanism was initially implemented in three pilot provinces (i.e., Fujian, Zhejiang, and Sichuan) in China in 2004 and was subsequently promoted to the remaining 29 provinces (with the exception of Xinjiang and Tibet) in July 2012.

The introduction of the new pricing mechanism would not only affect residents' electricity bills, but also lead to a change in their electricity consumption behaviours. Under the traditional low-price flat tariff, differences in consumption behaviours amongst different groups of residents is not significant. However, when it comes to the new IBT, it appears that the behaviours of resident groups differ greatly. For example, price-sensitive residents may reduce their electricity usage when approaching the upper bound of the current block to avoid entering the next-higher price block. Conversely, price-insensitive residents may give more attention to comfort and convenience without considering the extra expenditures caused by the IBT. Their electricity demands are less responsive to the change of electricity tariff. In addition, under the IBT, the responses of different residential groups to weather conditions and other external factors may not be the same. The different types of response to the new tariff, in turn, cause a change in electricity demands.

Load forecasting plays an essential role in modern power system planning, scheduling, operation, maintenance, and control, and it has received a great deal of attention from researchers [10, 28, 41, 44]. In practice, load forecasting can be classified into three types according to the forecasting time interval, including short-term, medium-term, and long-term load forecasting, and these types are applied to different application scenarios. Medium- and long-term forecasting predict load demand from a few weeks to several years in advance, and they are commonly used for the planning, maintenance, reconstruction, and development of

a power system that will satisfy future demands [40, 41, 46]. Short-term load forecasting predicts system load for a brief interval, from minutes to several days, and it is the most commonly and widely used of the three types. Short-term load forecasting is crucial to the day-to-day financial operation, scheduling, and load-shedding plans of power utilities, and in recent years it has become increasingly popular for dynamic demand modelling and real-time pricing for deregulated electricity markets [25, 28]. However, the introduction of the new IBT makes short-term load forecasting more complicated.

In an effort to remedy the pressing issues outlined above, this study proposes an innovative clustering-based load forecasting approach to support short-term load forecasting for residential electricity under the IBT in China. With the development of smart metering in China, the system for real-time electricity-usage collection and retrieval records more abundant and more accurate residential consumption data. This will provide the data necessary for capturing and analysing the consumption patterns of different resident groups under the IBT. The new approach firstly groups residents into different segments according to their consumption patterns by using the fuzzy C-means (FCM) clustering algorithm [6, 19], then builds load forecasting models for each segment, and finally aggregates the predicted usages of different clusters to derive the total usage. The utility and effectiveness of the proposed model is verified through a realistic dataset that contains the daily household-level consumption data of 533 households from April 2014 to February 2015. The results show that the proposed approach can not only achieve better prediction accuracy, but also provide better flexibility for hybrid modelling.

This novel study offers several theoretical and practical contributions. First, it introduces a new modelling approach. The new fundamental mechanism of the proposed model is different from both the conventional overall aggregation approach and the individual-level approach. The main advantages of the proposed approach are its capability to distinguish amongst different usage behaviours of customer groups and its capacity to enable the fore-

casting model that is appropriate to a given customer segment. Second, this study is the first to take the IBT-related attributes into account for the segmentation of residents. Such attributes offer new perspective to further explore and analyse the derived consumption behaviours under the IBT. Third, identifying the impact of a certain input variable/attribute is a challenging and difficult task in the extant load forecasting approaches. In particular, if the value of an attribute does not change within the given historical data (such as a price-related variable in the load forecasting), its impact will be implausible to be identified. An important value of the proposed approach both from the theoretical and application points of view is that, via the clustering of the different responses to this type of variables, their impacts are, to some extent, considered and addressed implicitly in the new forecasting model. Therefore, the clustering-based approach enhances the representation power and enables the flexibility of the new model to deal with those variables whose impact are difficult to be identified. Fourth, most of the extant research on load forecasting uses the aggregated-level load demand for prediction. However, this study offers a new method that demonstrates household-level daily consumption data can be used to produce more accurate prediction. Fifth, this study provides important practical implications by deep learning the identified consumption patterns, including support load demand forecasting, help to improve the design of the IBT, promote awareness and acceptance of the IBT, and it also has important potential application to market-based and priced-based demand side management.

The remainder of the paper is organised as follows. Section 2 briefly summaries the extant literature on the clustering of households and short-term load forecasting with the aim to highlight the research gap. After that, an innovative clustering-based load forecasting approach that enables the prediction of short-term load demand for a given area under the IBT is proposed. In Section 3, an empirical application of the proposed model is presented, and the feasibility and effectiveness of the proposed approach is also verified. The model comparisons and extensions are provided in Section 4, and Section 5 outlines the model's

implications and concludes the study.

2. Clustering-based short-term load forecasting approach

2.1. A brief literature review

The fundamental idea of the IBT is to identify different groups of users in the electricity market and then provide price discrimination for distinct groups, with the purpose of improving residential welfare, social equality, and efficiency, as well as alleviating energy pressure [37, 48, 52]. In order to address the short-term load forecasting problem under the IBT and reveal the research gap, a brief review of the extant literature on the clustering of households and short-term load forecasting is provided in this subsection.

2.1.1. Clustering of households

Intensive research efforts have been devoted to the segmentation of the electricity market. In general, the segmentation approaches can be divided into two streams: the attribute-oriented and consumption-pattern-oriented. The commonly used segmentation attributes include electricity usage, residential income, and temperature [1, 2, 29, 37, 51, 58] (see details in Table 1). With the introduction of smart meters in recent years, high-resolution time series consumption data (e.g., 15-minute, 30-minute, and 1-hour resolution) have become available. Research on household clustering has begun to change its focus from attribute-oriented factors (e.g., sociodemographic and building-related attributes) to consumption-pattern-oriented factors (load profiling). Table 1 lists representative segmentation studies that were performed on 15-minute, 30-minute, or hourly household load profiles.

Based on the load curves, some meaningful curve features/indices (e.g., off-peak factor, load factor, and modulation coefficient) were extracted for clustering. However, a load-profile-oriented approach to segmentation requires high-resolution (i.e., hourly or subhourly) smart meter data from residential consumers. This may not be widely available for regions

Table 1: Review of literature drawing on clustering of households.

Author(s)	Method(s)	Attributes	Type	Results
Albert and Maasoumy [1]	The proposed bi-section algorithm	304 variables such as sociodemographic and building-related customer characteristics	Attribute-oriented	Five clusters: high consumption, cost conscious, home improver, green advocate, and cultural drivers
Albert and Rajagopal [2]	K-medoids	Consumption characteristics: magnetite, duration, variability, and weather sensitivity	Attribute-oriented	Eight clusters
Hyland et al. [26]	OLS regression	Age, household income, employment status, education level, social status, gender, number of household members, number of bedrooms, and tenure	Attribute-oriented	Reveals the degree to which gross margin varies across customers with different characteristics
Lin et al. [37]	Cut on thresholds	Per capita annual disposable income, and per capita annual consumption expenditure	Attribute-oriented	Four clusters: poverty group, low income group, middle income group, and high income group
Zhang et al. [58]	Mlogit model	Gender, age, occupation, education, house income, and location	Attribute-oriented	Varying willingness to pay for green electricity across customers with different demographic variables
Benítez et al. [4]	Modified K-means algorithm	Hourly load profiles of Spanish residential customers and seasonality	Pattern-oriented	Ten distinct user clusters
Kwac et al. [29]	Adaptive K-means	Features extracted from normalised hourly load shape of a population of 220K residential consumers	Pattern-oriented	16-20 most frequent load shapes are identified
López et al. [30]	A Hopfiled-K-means algorithm	Hourly load profiles of 230 customers	Pattern-oriented	13 clusters
McLoughlin et al. [43]	K-means, K-medoids, and SOM	Hourly load profiles and dwelling, occupant and appliance characteristics of 3,941 customers	Pattern-oriented	Nine clusters, mapped to 10 electricity load profile classes
Räsänen et al. [49]	SOM + K-means/hierarchical clustering	Hourly measured electricity use profile of users in Finland	Pattern-oriented	19 distinct user clusters

or counties where smart meter implementation has not yet achieved scale. Further, it is evident in Table 1 that because load profiles often carry comprehensive features, especially when the time resolution is high, the number of detected clusters is relatively large (e.g., [4, 29, 30, 43, 49]). It is common to find more than ten detected clusters in extant studies. If the number of households is not large enough, some clusters will contain a very small portion of households. The large number of detected clusters makes it difficult to distinguish and analyse consumption characteristics, and it then becomes challenging to design and implement pricing discrimination and operation schema accordingly. In addition, to the best of the authors' knowledge, no extant study has taken the IBT-related attributes into consideration in the segmentation of the electricity market. The present study is the first to introduce IBT-related segmentation attributes.

2.1.2. Short-term load forecasting

The popular load forecasting methods can be classified into three broad categories: **conventional methods**, **computational intelligence techniques**, and **hybrid methods**. The main conventional methods, which are based on statistical models, are the time series approach and the multivariate regression approach. The time series approach is based on the hypothesis that future load demand is closely related to historical load values. Varieties of the autoregressive integrated moving average (ARIMA) model [7], which are representative of time-series-based methods, have been widely used in load forecasting (e.g., [15, 31, 57]). In multivariate regression, the predicted load is considered a linear combination of multiple explanatory variables, such as weather or seasonal variations, holiday activities, and demographic attributes. Examples can be found in the works of [18, 25].

Due to their mature and solid theory, conventional methods have achieved some promising results. However, such methods assume that the predicted load demand and the explanatory variables are linearly associated, which is difficult to achieve in reality. Therefore, computational intelligence techniques have been introduced to learn the complex and nonlin-

ear relationships in load forecasting, including fuzzy inference systems [56], artificial neural networks [28], Bayesian regression [54], causal models [45], and support vector machines [8]. Of these approaches, artificial neural networks represent the most popular method due to their attractive modelling features, such as the capacity to model any nonlinear function with good prediction accuracy, robustness in the presence of noisy data, and massive parallelism [3, 10, 28]. However, this approach suffers from poor interpretability and the overfitting problem. In addition, the selection of learning parameters and network structures, which have a positive impact on model performance, creates additional complications for end users.

To complement the drawbacks of different approaches, hybrid methods have become a promising research direction for load forecasting in recent years. These hybrid approaches can be classified into three types: combination (e.g., [9, 11, 28]), optimisation (e.g., [3, 10, 23, 39, 41]), and the integration of both (e.g., [34, 40, 57]). The first often applies several prediction models to the given problem and combines the prediction results derived from different prediction methods. The most commonly used aggregation method is a linear combination of base prediction models, devised by assigning equal or weighted vote scores. The second often selects a mother prediction model and then employs one or more other methods to optimise the learning parameters of the mother model with the aim of improving prediction accuracy. A commonly adopted approach, for example, is to combine the neural network method with other techniques to optimise the learning parameters and automate the selection of network structures [3, 10, 41]. The third type of approach is an integration of the first two types. Table 2 provides a summary of representative hybrid models for short-term load forecasting.

2.2. Overview of approach

In this study, when switching from the conventional low-price flat tariff to the new IBT, the consumption behaviours of users may change. Gaining a good understanding of these consumption patterns will enable more accurate load forecasting and better support the

Table 2: Review of literature drawing on hybrid approaches for short-term load forecasting.

Author(s)	Model(s)	Ensemble type	Data type	Data resolution	Approach
Khawaja et al. [28]	Boosted neural networks (BooNN)	Combination	Aggregation-level regional data from New England	1 h	BooNN consists of a combination of ANNs, and the final predicted result is the weighted sum of output from the individual trained ANNs.
Che and Wang [11]	Kernel-based support vector regression	Combination	Aggregation-level regional data from New South Wales	30 min; 1 h	A new method is proposed to select the optimal kernel functions for SVRs, and the overall predictions are derived by combining individual models through weighted votes.
Chahkoutahi and Khashei [9]	a novel DOPH based on MLP, ANFIS, and SARIMA	Combination	Aggregation-level regional data from New South Wales	30 min	The main idea of the proposed model is to simultaneously use advantages of base models (i.e., MLP, ANFIS, AND SARIMA) in modeling complex and ambiguous systems in a direct and optimal structure.
Chaturvedi et al. [10]	ANN, wavelet transform, adaptive genetic algorithm, and fuzzy system	Optimisation	Aggregation-level data from an institute in India	1 h	Apply the wavelet transform, adaptive genetic algorithm, and fuzzy system to optimise the learning parameters of ANN.
Mao et al. [41]	SOFNN and bilevel model	Optimisation	Aggregation-level regional data from Hebei Province, China	15 min; 30 min	Apply bilevel model to optimise the pretraining parameters of the SOFNN.
Liu et al. [39]	Empirical model decomposition (EMD), extended Kalman filter (EKF), ELM with Kernel, and particle swarm optimisation (PSO)	Optimisation	Aggregation-level data from a city of Zhejiang Province, China	1 h	EKF and ELM are adopted to predict different kinds of intrinsic model function, and PSO is used to optimise the model parameters.
Hanmandulu and Chauhan [23]	Wavelet fuzzy neural network (WFNN) and FNN employing Chaoquet integral (FNCI)	Optimisation	Aggregation-level regional data from India	1 h	In the WFNN, wavelets are employed to improve the modelling and learning capability of FNN. In FNCI, the use of the Choquet integral helps to speed up the reinforcement learning of FNN.
Bashir and Hawary [3]	Adaptive ANN, PSO, and wavelet transform	Optimisation	Aggregation-level regional data from New York	1 h	A PSO is employed to adjust the network's weights in the training phase of the ANNs. In addition, data are wavelet transformed during the pre-processing and then inserted into the ANNs to extract redundant information from the load curve.
Li et al. [34]	Wavelet transform, ELM, and PLSR	Integration of both	Aggregation-level regional data from New England and North American	1 h; 24 h	A wavelet-based ensemble strategy is proposed to generate different input features for the ELM-based predictors, and the PLSR is used to combine the outputs of individual predictors.

design of the IBT. To fill this research gap, this study proposes a new clustering-based approach for short-term load forecasting under the IBT. The proposed model is depicted in Figure 1, and its main steps are described as follows:

Figure 1: The proposed clustering-based load forecasting approach.

* **Data preparation:** The data collected from smart meters must be preprocessed, the main tasks include: 1) data transformation, 2) handling missing values, and 3) handling noisy data. In some cases, the smart meters capture the accumulated load values, such data should be transformed to calculate the daily usage data. Second, due to technique issues (e.g., smart meters occasionally receive abnormal signals), the missing values need to be handled. The commonly used methods include: deletion, manually filled, and automatically filled by using fixed values, mean/medium value, or inferred values. The next step is to handling the noisy and erroneous data, and it requires some expertise and careful examination of the given data. For example, the dataset may contains the load data for a residential public road lamp, and some apartments are not occupied within the observed time period. Such outliers need to be filtered out before passing for further analysis. Besides that, a data exploration can also be performed in this step to capture the basic load patterns, and the exploration will be helpful to determine the clustering attributes in the next step.

* **Selection of clustering attributes:** The clustering results are closely related to the selection of clustering attributes that reflect the segmentation criteria. The commonly used segmentation attributes are discussed in Section 2.1.1. In the proposed model, the cluster-

ing attributes are considered to better represent the electricity consumption behaviours of residents under the IBT. Hence, besides the traditional usage-related attributes, the IBT-related attributes are innovatively considered in the clustering analysis. In addition, It is well recognised that weather sensitivity varies across different household groups. The temperature attribute is often considered the clustering of consumers in the electricity market [2, 4]. Therefore, the temperature-related attributes are included to group households with similar consumption patterns into the same cluster. Note that, the proposed dimensions of attributes are general, and they can be used to guide the selection of customised clustering attributes for the given data.

* **Clustering analysis:** In this step, clustering algorithms will be employed to grouping households based on the selected clustering attributes and the characteristics of each identified segment will be analysed and discussed. For some clustering algorithms (e.g., K-means and Fuzzy C-means (FCM)), the number of clusters needs to be predefined. Determining the “right” number of clusters is crucial, because it controls the proper granularity of cluster analysis. In the literature, many validation metrics that consider different properties of clustering outcomes have been proposed [21]. However, it is well recognised that clustering is a kind of unsupervised learning technique, and there is no so-called “best” metric that can ensure optimal clustering results. In the proposed model, the number of clusters is determined by employing the aggregated results from several validation metrics. More specifically, several validation metrics are calculated using different numbers of clusters, respectively. To obviate dependence on measurement units, normalisation needs to be conducted on the absolute values of metrics. The larger normalised value indicates better performance for all metrics, and the normalised values are then summed up to represent an overall performance index for different numbers of clusters. By the end of this step, each household has an associated cluster label, and each household cluster has their own consumption patterns.

* **Data separation:** Assumed that N is the derived optimal number of clusters from the

last step, the processed dataset will then be separated into N sub-datasets, each of which will include households from the same cluster. Note that, this step distinguishes the proposed model from the extant hybrid models. As shown in Table 2, the hybrid models often build different forecasting models based on the same dataset, while different models are built upon different datasets in this study.

* **Forecasting models:** This step builds short-term forecasting model for each detected cluster, respectively. In other words, the proposed model allows each cluster to apply the most appropriate model to capture the consumption pattern. In each separated dataset, the daily usage of all household in the same cluster is initially aggregated, and the earlier days of load demands are used to train the forecasting model, while the later part is used for testing. To that end, the predicted usages of different clusters are aggregated to derive the total load demand. To verify the performance of the proposed model, the predicted load demands are compared against the actual load, so that load forecasting evaluation metrics (e.g., the mean absolute percentage error (MAPE)) can be reported.

In essence, the main feature and novelty of the proposed forecasting model is the geometrical (clustering-based) combination of several models to obtain a more accurate forecasting, and this is fundamentally different from traditional load forecasting approaches. From the data aggregation point of view, in general, extant approaches of load forecasting (see Table 2) can be divided into two types: **the overall aggregation approach** [3, 10, 11, 23, 28, 34, 39, 41, 54] and **the individual-level approach** [44]. The overall aggregation approach, a prevalent method in load forecasting, employs aggregation-level historical usage data to predict the total electricity usage of all residents within a certain area. All the electricity usages are aggregated in order to build an average consumption model. However, this approach ignores the diverse consumption behaviours of individuals and fails to reveal the consumption patterns of different resident groups. At the other extreme, the individual-level approach builds predictive models for each individual resident in

a certain area, and the total electricity usage is then derived by aggregating the individual usages [44]. Although this approach is useful for understanding individual consumption patterns, its load forecasting accuracy is usually poor due to the loss of mutual compensation of usage variations between customers. Further, when the number of included households is large, deriving the results is expensive and infeasible, and analysing and summarising the consumption characteristics is challenging. It also requires individual-level consumption data, which are relatively difficult to obtain. For example, due to the lack of realistic data, the proposed model in [44] was tested on a simulated dataset.

Having recognising these limitations, the proposed model, to some extent, provides an effective mean to filling the research gap. Compared to the overall aggregation approach, the new model avoids just using one forecasting model to cover different usage behaviours, which reduces prediction accuracy. Compared to the individual-level approach, the proposed model reduces the uncertainty and variation of individual usage behaviour via forecasting based on the similar behaviour groups and then improve the accuracy.

3. An empirical application of the proposed approach

3.1. The dataset

This section reports an empirical study conducted in order to verify the effectiveness and utility of the proposed approach. In this study, a realistic dataset that included the daily electricity usage data of 653 households under the IBT was collected. In contrast to extant studies, this dataset comprises 10 months (i.e., 11th April 2014 - 10th February 2015) of household-level daily usage data from an area in Fengze District, Quanzhou City, Fujian Province, China. According to the China Statistical Yearbook for regional economy (2014) and the China City Statistical Yearbook (2014), in Fengze District, the population was 556,000, gross regional product per capita (GRPPC) was 42,696 RMB, and per capita disposable income was 41,617 RMB in 2014. The dataset used in this study was collected

from a large housing estate in Fengze District, which is a representative residential area and includes various apartment types (e.g., one-bedroom, two-bedroom, three-bedroom, and four-bedroom apartments). Apartment size somehow reflects the income level of households; hence, the selected samples may support the investigation of the diversity of electricity consumption patterns. In addition, this housing estate, located in the central part of Quanzhou City, installed smart meters at a relatively mature stage, and it therefore has more complete usage data. The structure of the IBT employed in Fujian Province during 2014/5 is defined in Table 3. In this area, the starting date for the IBT cycle is the 11th of every month, and this study selected the data of the first 272 days (i.e., 11 April 2014 through 10 January 2015) for model training and selected the data of the remaining 34 days (i.e., 11 January 2015 through 10 February 2015) for testing.

Table 3: Structure of the IBT employed in Fujian Province during 2014/5.

Blocks	Monthly volume per household (kWh)	Price (CNY/kWh)
First	≤ 200	0.446
Second	201-400	0.5483
Third	> 400	0.7983

3.2. Data preprocessing

Because the raw dataset captures the accumulated load values, the first step is to calculate the daily usage data by using the current record minus the record of the previous day. More specifically, the daily usage of the i^{th} day (i.e., x_i) is calculated as $x_i = v_i - v_{i-1}$, where v_i is the load record of the i^{th} day. After the data transformation, the dataset includes the daily usage data of 653 households, and it consists of 199,858 consumption data points. The collected dataset contains some missing values, considering that the number of data samples is not very large, this study uses the Lagrange interpolation [5] to fill in the missing values. The dates and their corresponding load values are taken into consideration in interpolation. Note that, if a record has more than 30 missing values in succession, it is directly abandoned.

A closer examination of the dataset reveals that there exist some noisy and erroneous data points. Initially, if the average daily usage of a certain household is less than 0.5 kWh, these households are classified as outliers (e.g., road lamp) and are removed. In addition, if the number of zero values of a household's daily usage is greater than a threshold (i.e., 100 in this work), it is highly likely that this apartment is not occupied, and they are also eliminated. In the end, the cleaned dataset includes 533 households and consists of 163,098 consumption data points. An exploration of the dataset reveals that temperature plays a crucial role in electricity consumption; thus, the historical temperature data of Quanzhou within the observed time period is retrieved and included. The load patterns (including average daily loads and accumulative loads) and weather data are depicted in Figure 2.

Figure 2: Load patterns and weather information of the experimental dataset.

3.3. Clustering analysis and results

3.3.1. Selection of clustering attributes

In this study, besides traditional usage attributes, two IBT-related attributes are newly introduced into the clustering analysis. In addition, an examination of the collected dataset indicates that weather conditions affect electricity consumption through electrical appliances (e.g., heaters and air conditioners). Because the dataset was collected from the southern part of China, the summer months often feature high temperatures that require the intensive use of fans and air conditioners, which results in greater electricity consumption. By contrast, the winter months are relatively warm, and no significant increase in electricity consumption through the use of heating appliances is observed. Therefore, this study considers only consumption behaviours in the context of high temperatures. The selected clustering attributes, described below, are summarised in Table 4.

- Average daily load. This attribute directly reflects the electricity consumption of households. It is calculated by using the total consumption usage of a given household to divide the number of observation days. In addition, consumption usage is affected by the number of electrical appliances in a household, and this reflects the level of residential income to some extent [25, 37]. Households with similar income levels often have similar electricity consumption behaviours [12, 29, 37]; therefore, it makes sense to group them in the same segment.
- Second block ratio and third block ratio. These attributes refer to the ratio of the number of months that reaches the second block and the third block to the total number of months. They reflect the fluctuation of electricity consumption in the previous months and reveal consumer responses to the IBT. For instance, price-sensitive consumers tend to reduce their electricity usage when approaching the upper bound of the current block to avoid entering the next-higher price block; consequently, the derived

second and third block ratios will be small. On the other hand, a large third block ratio indicates that the consumer is either unaware of the IBT or cares more about comfort and convenience than the extra expenditures caused by entering the higher price block.

- High temperature sensitivity. This attribute refers to sensitivity to hot weather, and it is defined as the ratio of the average load on high-temperature days to the average daily load. In this study, temperatures are over 25°C are considered high. Larger values for this attribute indicates that consumers are more sensitive to high temperatures, because they tend to consume more electricity during summer months through the use of fans and air conditioners.

Table 4: List of selected clustering attributes.

Attributes	Definitions
Average daily load	Total load/number of observed days
Second block ratio	Number of months that reach the second block/number of total observed months
Third block ratio	Number of months that reach the third block/number of total observed months
High temperature sensitivity	Average load in high temperature days/average daily load

3.3.2. Determination of the number of clusters

After identifying the clustering attributes, their values need to be calculated. To eliminate the effects of measurement units and better protect data privacy, the *max – min* normalisation method is employed to transfer the derived values of clustering attributes into a range of $[0, 1]$. The normalised dataset is then passed to the following analysis.

This study employs the FCM algorithm [6, 19], one of the most prevalent clustering techniques, to cluster households based on their consumption behaviours. FCM extends the conventional K-means clustering algorithm by allowing the data point to belong to a cluster with a certain degree. Given an input dataset $I = \{x_i | i = 1, 2, \dots, N\}$ with N data points,

the basic idea of FCM is to produce optimal c -partitions of I by minimising the within-cluster sum of squares. In this study, the Euclidean distance [21, 24] is used to measure the difference of consumption behaviours between two consumers. The details of an FCM implementation can be found in [6, 20, 22].

In FCM, the number of clusters needs to be predefined. In this study, three widely used measures, the mixed pseudo F-statistic (mixed-F) [6], the Davies-Bouldin (DB) index [14], and the mean index adequacy (MIA) [13], are employed to determine the number of clusters. More specifically, the mixed-F measures the degree of the compactness and the separation of generated clusters, the DB index concurrently considers the degree of within-cluster scatter to the between-cluster separation, and the MIA index calculates the average distance between the cluster centroids and data points of a given cluster. For the mixed-F, a higher value of the index indicates better performance, whereas for the DB and MIA indexes, a smaller value of the index indicates better performance.

These three measures are calculated using different values of c (i.e., the number of clusters), respectively, and the obtained results are reported in Table 5. After that, the *max-min* normalisation method is employed to transform the absolute values of metrics to fall within a range of $[0, 1]$. It is shown in Table 5 that when $c = 5$, the clustering achieves optimal overall performance (see the last column in Table 5). Therefore, the households are segmented into five clusters for further analysis in this study.

Table 5: Clustering validation index.

Number of clusters	Mixed-F		DB index		MIA index		Overall
	Absolute	Normalised	Absolute	Normalised	Absolute	Normalised	
3	101.6	0.0	1.0838	0.1908	0.2897	0.0	0.1908
4	197.8	0.5833	0.9577	0.7727	0.2424	0.4407	1.7967
5	266.5	1.0	0.9085	1.0	0.2043	0.7958	2.7958
6	211.8	0.6683	1.1251	0.0	0.1940	0.8923	1.5606
7	217.9	0.7053	1.0854	0.1836	0.1824	1.0	1.8889

3.3.3. Results of clustering

As a result of applying FCM, the 553 households are grouped into five segments, referred to here as $Cluster_1$, $Cluster_2$, $Cluster_3$, $Cluster_4$, and $Cluster_5$. The clustering results and the details of cluster centroids are summarised in Table 6. To better illustrate the consumption behaviours of different clusters, Figure 3 depicts the daily load of individual households in different clusters from 11 April 2014 through 10 February 2015. The characteristics of the five detected clusters are summarised as follows:

- $Cluster_1$ (**low-demand users and insensitivity to high temperature**) accounts for 14.44% of the population. The households in $Cluster_1$ consume the least electricity, and the daily consumption is very stable and often below 10 kWh per day. The electricity usage of this cluster stays mainly in Block 1, which is designed to cover essential living needs, and it rarely enters Blocks 2 and 3. In addition, the users in this cluster are insensitive to high temperature. As shown in Figure 3, the users do not consume more electricity in the summer.
- $Cluster_2$ (**ordinary users and sensitivity to high temperature**) constitutes 21.95% of the population. The electricity usage of this cluster stays mainly in Blocks 1 and 2. However, according to Figure 3, users in this cluster have the highest temperature sensitivity (i.e., 0.6272). Therefore, they consume more electricity in the summer, and the electricity usage reaches Block 3 in certain months. It seems that the users in this cluster are more concerned about the comfort and quality of life than the cost of electricity. They do not mind paying a higher price to use the air conditioning in the summer, even though the price reaches Block 3.
- $Cluster_3$ (**ordinary users and sensitivity to the IBT**) accounts for 16.51% of the population. The electricity usage of this cluster stays mainly in Block 2 with a high second block ratio (i.e., 0.7951), and the users are not very sensitive to high

Figure 3: Electricity load curves of different clusters.

temperature. Electricity consumption in summer is only slightly higher than in other seasons. Interestingly, although the average daily load of this cluster is higher than that of $Cluster_2$, electricity usage rarely reaches Block 3. A possible reason for this is that the users in this cluster are more sensitive to the IBT and tend to reduce their electricity usage when approaching the upper bound of Block 2 to avoid entering Block 3, which results in a higher price.

- $Cluster_4$ (**high-demand consumers and sensitivity to high temperature**) accounts for 36.4% of the population. The electricity usage of this cluster is almost equally distributed across Blocks 2 and 3, and this cluster consists of high-demand consumers. In this cluster, high-temperature sensitivity ranks in second place amongst all clusters. Thus, there is a significant consumption increase in summer.
- $Cluster_5$ (**luxury consumers**) constitutes 10.70% of the population. The average

daily load gradually increases from $Cluster_1$ to $Cluster_5$, and the users in this cluster consume the most electricity. The electricity usage of this cluster is stable and stays mainly in Block 3. Therefore, this cluster consists of luxury consumers who are price insensitive and consume as much electricity as they need without considering the extra expenditures caused by the IBT. These luxury users naturally consume more electricity in hot weather.

According to the above analysis, the IBT has little impact on $Cluster_1$ and $Cluster_5$, as their demands are quite stable and are less responsive to the change of electricity tariff. Under the current IBT, the usage of $Cluster_1$ stays mainly in Block 1, whereas the usage of $Cluster_5$ remains mainly in Block 3. The consumers in $Cluster_2$ and $Cluster_5$ are relatively less sensitive to cost; they tend to pay more attention to comfort and convenience without considering the extra expenditures caused by the IBT. By contrast, the IBT seems to have some impact on $Cluster_3$, because although the average daily load of this cluster is higher than that of $Cluster_2$, the users have made some efforts to avoid entering the higher price block. For $Cluster_2$ and $Cluster_4$, the impact of the current IBT is not significant. However, the IBT could potentially have a stronger influences on these clusters, because their electricity usages are distributed to two neighbouring blocks; it would therefore be possible to switch the usage from one block to another, if consumers intend to take some action.

3.4. Short-term load forecasting

Based on the clustering results, the collected households are grouped into five segments that have their own characteristics. Five load forecasting models for each segment are then built, respectively. This study employs the Self-organising fuzzy neural networks (SOFNN) prediction algorithm [32] for this task, and each forecasting model differs from each other in terms of network structure and parameters. SOFNN was initially proposed by Leng et al. in 2004 [32], and it is a simple and effective approach to generating a fuzzy neural

Table 6: Summary of clustering results.

Cluster	Average daily load	Second block ratio	Third block ratio	Temperature sensitivity	Characteristics	Number of households	Percentage
<i>Cluster</i> ₁	0.0603	0.0351	0.0072	0.3588	Low-demand users and insensitivity to high temperature. Users in this cluster consume the least electricity. Demand is stable and very low, even in summer.	77	14.44%
<i>Cluster</i> ₂	0.1732	0.3750	0.1297	0.6272	Ordinary users and sensitivity to high temperature. In this cluster, usage stays mainly in Blocks 1 and 2, but users are very sensitive to high temperature. Demand increases sharply in summer.	117	21.95%
<i>Cluster</i> ₃	0.1994	0.7951	0.0911	0.4362	Ordinary users and sensitivity to the IBT. Usage stays mainly in Block 2 and rare reaches Block 3, even in summer.	88	16.51%
<i>Cluster</i> ₄	0.3146	0.4995	0.4655	0.6020	High-demand consumers and sensitivity to high temperature. Usage is almost equally distributed across Blocks 2 and 3, and users consume considerably more in summer.	194	36.40%
<i>Cluster</i> ₅	0.4530	0.1031	0.8848	0.5283	Luxury consumers. Usage is stable and stays mainly in Block 3. Users consume as much as they want without considering the extra costs caused by the IBT.	57	10.70%

model with high approximation accuracy and a compact network structure. The reason of choosing SOFNN algorithm is that the proposed model requires to build several forecasting models for different clusters, and SOFNN can self-organises its own neurons in the learning process, in which both the learning parameters and network structure are automatically updated. This helps to achieve better prediction accuracy and relieve the burden on end users. Its superiority over other models (e.g., ANFIS and FNN) has been demonstrated in [33, 41]. An SOFNN consists of five layers: the input layer, ellipsoidal basis function (EBF) layer, normalised layer, weighted layer, and output layer. The details of an SOFNN implementation can be found in [32, 41].

The input attributes of the SOFNN consist of two parts: the historical load data and the temperature data. More specifically, in order to predict the load of the i^{th} day, the previous seven days' load data, together with the predicted temperature of the i^{th} day (i.e., in total 8 attributes), are fed into the SOFNN for modelling. This process continues until the end of the experimental time period. To determine the most appropriate learning parameters for the SOFNN, the trial-and-error method is employed in pre-experiments for parameter selection. The initial parameters are defined by following the general guidance provided in [32] and referring the values that used in previous studies (e.g., [32, 41]). For example, it is suggested in [32] that the distance threshold (kd) can be defined as the 10-20% of the input data range, and an appropriate value of the initial width (σ_0) is about two times of the smallest value of the input distance threshold. The smaller value of the error tolerance (δ) and the expected training RMSE ($krmse$) can achieve better prediction performance, but result in more complex network structure and more expensive computational costs. These initial parameters are then adjusted by trail-and-error tuning, a number of parameter value combinations are tested. In other words, the values of these parameters are determined by analysing the performance of the SOFNN vis-à-vis different parameter setting scenarios. For example, the values of δ and σ_0 are gradually increased starting from the value of 0.01 to

the value of 1.0. Some typical values of $krmse$ that are often used in the literature, such as 0.01, 0.05, 0.1 and 0.2, are employed in the experiment. As a result, the following optimal learning parameters are obtained: $\delta = 0.04$, $\sigma_0 = 0.01$, $krmse = 0.01$, and $kd(i) = 0.1$ (i.e., $i = 1 \cdots 7$).

The forecasting accuracy is reported in mean absolute percentage error (MAPE) and maximal error (ME), which are commonly used in load forecasting [9, 11, 28, 41], and the measures are defined as

$$MAPE = \frac{\sum_{i=1}^N \left| \frac{y_i - \tilde{y}_i}{y_i} \right|}{N} \times 100\%, i = 1, \cdots N \quad (1)$$

$$ME = \max \left| \frac{y_i - \tilde{y}_i}{y_i} \right| \times 100\%, \quad (2)$$

where N is the number of testing days (i.e., $N = 34$ herein), and y_i and $\tilde{y}_i$ are the real and predicted values of the electrical load of the i^{th} training day, respectively. The obtained prediction accuracies of different clusters are shown in Table 7.

Table 7: Prediction performances using the SOFNN.

	<i>Cluster</i> ₁	<i>Cluster</i> ₂	<i>Cluster</i> ₃	<i>Cluster</i> ₄	<i>Cluster</i> ₅	Total
MAPE	8.93%	4.02%	4.33%	3.45%	3.45%	2.59%
ME	26.99%	11.27%	15.99%	13.84%	10.85%	9.11%

As shown in Table 7, with the exception of *Cluster*₁, the forecasting accuracy of the other clusters is satisfactory. It is possible that *Cluster*₁ includes some abnormal consumers, such as residents who are often away from home for business trips or holidays. The clustering-based approach groups such consumers into the same cluster (i.e., *Cluster*₁). Because the consumer behaviours of these users are quite random and vary, they are difficult to predict using one forecasting model. On the other hand, it is challenging to build the predictive model individually. Thus, this cluster exhibits poor prediction performance. However, because the number of households in this cluster is not very large (it accounts for only

14.44% of the population) and the electricity usage is relatively low, this cluster’s poor performance does not significantly affect the aggregated forecasting accuracy.

To further validate the effectiveness of the clustering-based approach (denoted as C-SOFNN), its performance is also compared against the conventional overall aggregation model (denoted as O-SOFNN). Note that, due to the expensive computational cost, the individual-level approach is not included for comparison. Using the same dataset and observation window (i.e., seven days), the forecasting performance of these two approaches is depicted in Figure 4.

Figure 4: Daily prediction accuracy using overall and clustering-based models in the case of SOFNN.

The overall aggregation model directly uses the past seven days’ load and the temperature data to predict the load of the eighth day without conducting the clustering analysis. For the C-SOFNN, the predicted load of individual households is aggregated to derive the total load, and the forecasting accuracy is reported in both MAPE and ME (see Table 7). The performance of the conventional O-SOFNN is derived as follows: $\text{MAPE} = 2.74\%$ and $\text{ME} = 12.79\%$. This figure indicates that the C-SOFNN outperforms the O-SOFNN on short-term load forecasting by achieving better MAPE and ME. This is because, rather than

employing one general model, the clustering-based approach is more customised, because it selects the more appropriate forecasting model for different household clusters based on their consumption patterns. In addition, on the 17th day (i.e., 25 January, 2015), the prediction accuracy of both models drops sharply to below 90%. The historical data indicate that a regional power failure caused a great reduction of electricity usage on that day. Both models failed to identify and respond to this emergency, so the overall forecasting accuracy is affected to some extent. It is worth noting that in this study, short-term load forecasting is conducted due to the availability of the dataset. However, the proposed approach is general, and it can be readily applied to medium- and long-term load forecasting if longer-term consumption data are available.

4. Model comparisons and extensions

As Section 3.4 shows, the clustering-based approach outperforms the conventional overall aggregation model when using the SOFNN for short-term load prediction. Whether this superiority exists only in the case of SOFNN is a valid question. To address this question, this section applies other widely used prediction algorithms to both the proposed clustering-based approach and the overall aggregation approach. It is interesting to investigate and compare the performance of models using different prediction algorithms. In this study, the ARIMA [8], support vector regression (SVR) [17], and particle swarm optimisation (PSO) [27, 53] are selected for model comparisons due to their maturity and popularity. The same training and testing datasets employed in Section 3.4 are used for these three models for comparison purposes. The obtained results are reported in Figure 5 and Table 8, where C-ARIMA stands for the clustering-based approach by using the ARIMA, and O-ARIMA stands for the overall aggregation approach by using the ARIMA. This notation also applies to the SOFNN, SVR, and PSO algorithms.

The obtained results indicate that the proposed clustering-based approach generally

Figure 5: Daily prediction accuracy in MAPE using different models.

produces better prediction performance (i.e., it has lower MAPEs) than the conventional approach, and this is not dependent on a particular prediction algorithm. The performance of the C-SVR and that of the O-SVR are quite similar. This may be the case because the overall aggregation approach actually averages the load demands of all consumers in the dataset, the prediction errors then can be neutralised when calculating the overall MAPE of daily loads. Hence, the performance of the overall aggregation approach can be improved.

Table 8: Prediction performances using different models.

	C-SOFNN	O-SOFNN	C-ARIMA	O-ARIMA	C-SVR	O-SVR	C-PSO	O-PSO
MAPE	2.59%	2.74%	3.32%	3.82%	2.41%	2.45%	2.84%	3.00%
ME	9.11%	12.79%	12.10%	14.42%	10.32%	11.44%	10.21%	11.10%

For the clustering-based approach, the above analysis applies one prediction algorithm to all the detected clusters of the dataset at a time. Table 9 reports the obtained MAPEs for different clusters by using different prediction algorithms. Even for the same algorithm,

prediction performance varies across different clusters. There is no algorithm that outperforms the others in all clusters. For example, the SOFNN performs best in *Cluster*₅ by achieving the lowest MAPE, whereas it performs worst in *Cluster*₁ and *Cluster*₃. These results clearly reflect that the applicabilities and characteristics of different algorithms can be quite different.

Table 9: Prediction performances of applying different prediction algorithms to different clusters.

Models	<i>Cluster</i> ₁	<i>Cluster</i> ₂	<i>Cluster</i> ₃	<i>Cluster</i> ₄	<i>Cluster</i> ₅	Total
C-SOFNN	8.93%	4.02%	4.33%	3.45%	3.45%	2.59%
C-ARIMA	6.10%	4.24%	3.68%	4.73%	3.65%	3.32%
C-SVR	6.23%	3.70%	3.41%	3.02%	3.53%	2.41%
C-PSO	6.29%	4.45%	3.78%	4.07%	3.74%	2.84%
Winner	C-ARIMA	C-SVR	C-SVR	C-SVR	C-SOFNN	2.28%

In view of this fact, it would be beneficial to improve the prediction performance of short-term load forecasting by choosing the most appropriate algorithm for different consumer clusters. In essence, the proposed approach makes such hybrid methods plausible. As shown in Table 9, given a certain cluster, all available prediction algorithms will be exhaustively tested, and the winner algorithm will then be selected to model the load demand of the cluster. Each of the detected clusters are modelled respectively, and the obtained results are then aggregated to derive the combined prediction accuracy. When the new hybrid model is used, the derived MAPE is 2.28%, and the actual and predicted daily loads are depicted in Figure 6. It is shown that the prediction performance in terms of MAPE represents an improvement over using a single prediction algorithm for all clusters (see Table 9).

To further validate the superiority of the proposed hybrid model, the clustering-based approach is compared against several other conventional ensemble models. A conventional ensemble combines a series of prediction algorithms with the aim of creating an improved composite prediction model [11, 21]. The ensemble returns the overall prediction results based on the aggregations of the base predictors. More specifically, each base predictor's

Figure 6: Daily prediction accuracy in MAPE using hybrid models.

vote can be assigned an equal or weighted score, and the overall prediction result is then a linear combination of the results of base predictors. This study employs both the equal and weighted aggregation methods to derive the overall ensemble results. The weight of the i^{th} predictor's vote is defined as

$$w_i = \frac{e_i^{-1}}{\sum_{i=1}^k e_i^{-1}}, i = 1, 2, \dots, k, \quad (3)$$

where e_i is the prediction error of the i^{th} predictor and k is the number of hybrid prediction algorithms [11]. The weighting mechanism ensures that the predictor with better performance contributes more to the overall results. The obtained results of using different combinations of predictors in terms of MAPE are shown in Table 10, and it can be concluded that the proposed method outperforms the conventional ensemble models.

Table 10: Prediction performance of different hybrid models.

Hybrid models	Equal vote	Weighted vote
SOFNN, ARIMA, SVR	2.81%	2.73%
ARIMA, SVR, PSO	2.89%	2.81%
SOFNN, SVR, PSO	2.66%	2.65%
SOFNN, ARIMA, PSO	2.94%	2.89%
SOFNN, ARIMA, SVR, PSO	2.81%	2.76%
Proposed model	2.28%	

5. Implications and Conclusions

5.1. Implications

The proposed model has several practical implications. First, from the operational perspective, the predicted load demand may help power companies to provide a reliable and high-quality electricity supply as well as to establish appropriate schedules of operations and maintenance within a certain area. Normally, a certain area can be divided into several electricity distribution regions, and the power company provides a set of distribution transformers for each region. Rather than predicting the load demand for the whole area, the proposed model provides a more flexible modelling approach that enables the capture of consumption patterns of different distribution regions. As a result, the power company can accordingly design and implement customised operational strategies to ensure high-quality electricity supply services. For example, if the predicted load demand of a region is high, the power company could consider enlarging the transformation capacity in the near future.

Second, the obtained results can be used to analyse and improve the design of the IBT. The current IBT employs mainly the static analytical method for tariff design. For instance, according to the guidance on the implementation of the IBT for residential electricity consumption issued in [42], Block 1 must satisfy the basic electricity demand by covering 80% of the average monthly electricity consumption volume for a certain area, whereas Block 2 must cover 95%. In other words, 80% of the households in a certain area should be covered

in Block 1 in terms of their average monthly electricity consumption volume, whereas 95% of the households should be covered in Block 2. In addition, the price charged in Block 3 should not exceed 1.5 times the price in Block 2. According to this study's examination of the breakdown of the average monthly electricity consumption volume by tariff block, only 21.39% households are covered in Block 1 and 64.17% households are covered in Block 2. The detailed monthly breakdowns for the observed time period are shown in Table 11. The obtained results show that the current IBT in Fujian Province does not meet the initial requirements and that its effects are not very significant in Quanzhou City. The volume of electricity consumption for Block 1 and for Block 2 seems too low, because 35.83% of households will reach Block 3, and this ratio becomes much higher in the summer (see Table 11). Another possible reason for this is that the prices for Block 2 and Block 3 are not high enough to encouraging energy saving, as a result that many customers are insensitive to the extra costs of high consumption. These observations should be taken into consideration when the IBT design is revised. Designing the IBT is a complicated task, because several parameters need to be identified, including the number of blocks, and the volume and price for each block. An accurate demand forecasting approach plays an essential role in identifying such parameters, and these form the basis of the design of a reasonable IBT.

Table 11: Breakdowns of monthly electricity consumption volume by tariff blocks.

	2014.04	2014.05	2014.06	2014.07	2014.08	2014.09	2014.10	2014.11	2014.12	2015.01
Block 1	41.09%	33.40%	18.39%	15.76%	16.89%	20.83%	39.77%	42.03%	32.27%	29.46%
Block 2	49.53%	48.78%	32.08%	21.39%	24.58%	36.96%	51.97%	48.41%	46.53%	46.72%
Block 3	9.38%	17.82%	49.53%	62.85%	58.54%	42.21%	8.26%	9.57%	21.20%	23.83%

Third, the obtained results can be used to promote awareness and acceptance of the IBT. The clustering results obtained from this study have helped not only to identify households that are sensitive to the IBT, but also to identify target households that may be strongly influenced by the IBT. For example, the clustering results indicate that households in *Cluster*₃ exhibit more sensitivity to the IBT. In addition, the electricity usage of

households in $Cluster_2$ and in $Cluster_4$ is often distributed across two neighbouring blocks. These households have more of an opportunity to take certain actions in an effort to reduce electricity use and avoid entering the higher block. As a result, more promotional efforts should target $Cluster_2$ and $Cluster_4$.

Fourth, the proposed approach has important potential applications to market-based and price-based demand response or demand side management using day-ahead pricing, real-time pricing, critical-peak pricing, and time-of-use. In general, the short-term load/demand under flat prices has been relatively stable, and customers' consumption behaviours are predominantly affected by whether conditions and daily life patterns, whereas the impacts of other factors are limited. Therefore, the overall aggregation forecasting approaches have been shown to be effective and reliable when weather forecasting data are used in load forecasting. However, once market-based and price-based demand response management schemes based on smart meters are introduced into the market on a large scale, price impact will have a much more significant affect on customers' consumption behaviours than weather conditions, and different response patterns across customer groups will be unavoidable. Consequently, the conventional overall aggregation forecasting approaches may no longer be able to produce accurate short-term forecasting, even when taking prices as the model variables. The proposed model can overcome these weaknesses of the aggregated approach, and it has the advantages of great simplicity in computing and a high degree of accuracy in comparison with individual-level approach to load forecasting.

5.2. Conclusions

In China, much effort has been devoted to the reform of the residential electricity tariff in order to improve the efficiency of resource allocation, cultivate electricity saving, and reducing emissions [38, 52]. The different responses to the new pricing mechanism (i.e., the IBT) will change the consumption behaviours of residents, and such dynamic changes of load demands make load forecasting more complicated. To support this task, an innovative

clustering-based approach is proposed in this study to better capture the diverse consumption patterns under the IBT and increase the overall forecasting accuracy. The applicability and effectiveness of the proposed model are demonstrated through an empirical analysis of a realistic dataset in Quanzhou City.

The FCM clustering technique is used to classify households into five clusters that represent distinct consumption patterns, including ‘low-demand users and incentive to high temperature’, ‘ordinary users and sensitivity to high temperature’, ‘ordinary users and sensitivity to the IBT’, ‘high-demand consumers and sensitivity to high temperature’, and ‘luxury consumers’. The prediction performance of the proposed model is further confirmed by employing four prediction algorithms (i.e., SOFNN, ARIMA, SVR, and PSO) for short-term load forecasting. The new approach consistently achieves better prediction permanence in terms of MAPE than the conventional aggregation approach in all three cases. Moreover, the new approach allows for the development of a hybrid model that can employ different prediction algorithms for different clusters. A further improvement in terms of MAPE as a result of using this hybrid model is also evident in the comparison with other ensemble models. Accurate prediction performance would have a positive effect on improving the IBT design, ensuring a high-quality electricity supply, reducing operational costs, and establishing reasonable schedules of maintenance.

Although the current study offers some promising results, it has some limitations, and much progress can be made through further investigation. First, the data used in this study has some limitations. For example, the power company provides only load demand data, customer IDs, and addresses. It is difficult to link the demand data with customer-related attributes, such as income, age, house size, and number of family members. More efforts should be made to collect such demographic attributes, and then to apply more customer-related attributes in the clustering of households. In addition, due to the availability of the data, this study employs household-level daily consumption data for consumption analysis

and load forecasting. Although the prediction accuracy is satisfactory when compared to the high-resolution time series data, some detailed consumption behaviours are missed, such as electricity consumption in peak and off-peak time periods. Second, as pointed out in [16, 51], substantial differences in terms of consumption behaviour may exist within urban and rural areas. However, all the households in this experimental dataset are in an urban area, and their inhabitants may have similar lifestyles and habits. The diversity of the dataset needs to be further exploited by including households from different areas (e.g., rural areas and villages). As a result, more types of consumption patterns can be identified and compared through clustering analysis. The accurate clustering of households would provide great support for load forecasting. Third, the learning parameters of SOFNN are currently determined by the trail-and-error method, they can be automatically defined by using optimisation models (e.g., genetic algorithm) in future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 71301133, 91746103 and 71572166), the Fundamental Research Funds for the Central Universities (Project No. 20720161044), and Fuzhou Humanities and Social Science Project (No. 2017FZA05).

References

- [1] A. Albert and M. Maasoumy. Predictive segmentation of energy consumers. *Applied Energy*, 177:435 – 448, 2016.
- [2] A. Albert and R. Rajagopal. Smart meter driven segmentation: What your consumption says about you. *IEEE Transactions on power systems*, 28(4):4019 – 4030, 2013.
- [3] Z. A. Bashir and M. E. El-Hawary. Applying wavelets to short-term load forecasting using PSO-based neural networks. *IEEE Transactions on Power Systems*, 24(1):20–27, 2009.

- [4] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado. Dynamic clustering segmentation applied to load profiles of energy consumption from spanish customers. *International Journal of Electrical Power & Energy Systems*, 55:437 – 448, 2014.
- [5] J. Berrut and L. N. Trefethen. Barycentric lagrange interpolation. *Siam Review*, 2004.
- [6] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer, 1981.
- [7] G. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1970.
- [8] E. Ceperic, V. Ceperic, and A. Baric. A strategy for short-term load forecasting by support vector regression machines. *IEEE Transactions on Power Systems*, 28(4):4356–4364, 2013.
- [9] F. Chahkoutahi and M. Khashei. A seasonal direct optimal hybrid model of computational intelligence and soft computing techniques for electricity load forecasting. *Energy*, 140:988 – 1004, 2017.
- [10] D. Chaturvedi, A. Sinha, and O. Malik. Short term load forecast using fuzzy logic and wavelet transform integrated generalized neural network. *International Journal of Electrical Power & Energy Systems*, 67:230 – 237, 2015.
- [11] J. Che and J. Wang. Short-term load forecasting using a kernel-based support vector regression combination model. *Applied Energy*, 132:602 – 609, 2014.
- [12] G. Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68 – 80, 2012.
- [13] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *IEEE Transactions on power systems*, 18(1):381–387, 2003.
- [14] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, Feb. 1979.
- [15] E. M. de Oliveira and F. L. C. Oliveira. Forecasting mid-long term electric energy consumption through bagging arima and exponential smoothing methods. *Energy*, 144:776 – 788, 2018.
- [16] X.-Y. Dong and Y. Hao. Would income inequality affect electricity consumption? Evidence from China. *Energy*, 142:215 – 227, 2018.
- [17] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machine. *Advances in Neural Information Processing Systems*, 9:1155 – 161, 1996.
- [18] G. Dudek. Pattern-based local linear regression models for short-term load forecasting. *Electric Power Systems Research*, 130:139 – 147, 2016.
- [19] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated

- clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.
- [20] X. Fu, X. Chen, Y.-T. Shi, I. Bose, and S. Cai. User segmentation for retention management in online social games. *Decision Support Systems*, 101:51 – 68, 2017.
 - [21] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Morgan Kaufmann, 2011. Third Edition.
 - [22] S. Han, Y. Ye, X. Fu, and Z. Chen. Category role aided market segmentation approach to convenience store chain category management. *Decision Support Systems*, 57:296 – 308, 2014.
 - [23] M. Hanmandulu and B. Chauhan. Load forecasting using hybrid models. *IEEE Transactions on Power Systems*, 26(1):20–29, 2011.
 - [24] J. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.
 - [25] M.-F. Hung and T.-H. Huang. Dynamic demand for residential electricity in Taiwan under seasonality and increasing-block pricing. *Energy Economics*, 48:168 – 177, 2015.
 - [26] M. Hyland, E. Leahy, and R. S. Tol. The potential for segmentation of the retail market for electricity in Ireland. *Energy Policy*, 61:349 – 359, 2013.
 - [27] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the 2012 IEEE Conference on Neural Networks*, pages 1942–1948, Piscataway,USA, 1995.
 - [28] A. Khwaja, X. Zhang, A. Anpalagan, and B. Venkatesh. Boosted neural networks for improved short-term electric load forecasting. *Electric Power Systems Research*, 143:431 – 437, 2017.
 - [29] J. Kwac, J. Flora, and R. Rajagopal. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1):420–430, 2014.
 - [30] J. J. López, J. A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, and J. E. Ruiz. Hopfield-K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems Research*, 81(2):716 – 724, 2011.
 - [31] C.-M. Lee and C.-N. Ko. Short-term load forecasting using lifting scheme and ARIMA models. *Expert Systems with Applications*, 38(5):5902 – 5911, 2011.
 - [32] G. Leng, G. Prasad, and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. *Neural Networks*, 17(10):1477 – 1493, 2004.
 - [33] G. Leng, X.-J. Zeng, and J. A. Keane. An improved approach of self-organising fuzzy neural network based on similarity measures. *Evolving Systems*, 3(1):19–30, 2012.
 - [34] S. Li, L. Goal, and P. Wang. An ensemble approach for short-term load forecasting by extreme learning

- machine. *Applied Energy*, 170:22 – 29, 2016.
- [35] B. Lin and Z. Jiang. Targeted electricity-subsidy contributes to equity and efficiency (in Chinese). *Journal of Financial Research*, 11:48 – 58, 2009.
 - [36] B. Lin and Z. Jiang. Estimates of energy subsidies in China and impact of energy subsidy reform. *Energy Economics*, 33(2):273 – 283, 2011.
 - [37] B. Lin and Z. Jiang. Designation and influence of household increasing block electricity tariffs in China. *Energy Policy*, 42:164 – 173, 2012.
 - [38] B. Lin and X. Liu. Electricity tariff reform and rebound effect of residential electricity consumption in China. *Energy*, 59:240 – 247, 2013.
 - [39] N. Liu, Q. Tang, J. Zhang, W. Fan, and J. Liu. A hybrid forecasting model with parameter optimization for short-term load forecasting of micro-grids. *Applied Energy*, 129:336 – 345, 2014.
 - [40] Y. Liu, W. Wang, and N. Ghadimi. Electricity load forecasting by an improved forecast engine for building level consumers. *Energy*, 139:18 – 30, 2017.
 - [41] H. Mao, X.-J. Zeng, G. Leng, Y.-J. Zhai, and J. Keane. Short-term and midterm load forecasting using a bilevel optimization model. *IEEE Transactions on Power Systems*, 24(2):1080–1090, 2009.
 - [42] National Development and Reform Commission. Guidance on the implementation of the increasing-block pricing tariffs for the residential electricity consumption, 2011.
 - [43] F. McLoughlin, A. Duffy, and M. Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, 141:190 – 199, 2015.
 - [44] F.-L. Meng and X.-J. Zeng. A profit maximisation approach to demand response management with customers behaviour learning in smart grid. *IEEE Transactions on Smart Grid*, pages 1 – 14, 2015.
 - [45] J. Moral-Carcedo and J. Pérez-García. Integrating long-term economic scenarios into peak load forecasting: An application to Spain. *Energy*, 140:682 – 695, 2017.
 - [46] J. Nowotarski, B. Liu, R. Weron, and T. Hong. Improving short term load forecast accuracy via combining sister forecasts. *Energy*, 98:40 – 49, 2016.
 - [47] S. Okajima and H. Okajima. Estimation of Japanese price elasticities of residential electricity demand, 1990–2007. *Energy Economics*, 40:433 – 440, 2013.
 - [48] A. Prasanna, J. Mahmoodi, T. Brosch, and M. K. Patel. Recent experiences with tariffs for saving electricity in households. *Energy Policy*, 115:514 – 522, 2018.
 - [49] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, and M. Kolehmainen. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity

- use data. *Applied Energy*, 87:3538 – 3545, 2010.
- [50] P. C. Reiss and M. W. White. Household electricity demand revisited. *Review of Economic Studies*, 72(3):853–883, 2005.
 - [51] S. Silva, I. Soares, and C. Pinho. Electricity residential demand elasticities: Urban versus rural areas in Portugal. *Energy*, 144:627 – 632, 2018.
 - [52] C. Sun and B. Lin. Reforming residential electricity tariff in China: Block tariffs pricing approach. *Energy Policy*, 60:741 – 752, 2013.
 - [53] J. Sun, B. Feng, and W. Xu. Particle swarm optimization with particles having quantum behavior. In *Congress on Evolutionary Computation*, pages 325–331, Portland,USA, 2004.
 - [54] S. Wang, X. Sun, and U. Lall. A hierarchical bayesian regression model for predicting summer residential electricity demand across the U.S.A.. *Energy*, 140:601 – 611, 2017.
 - [55] Z. Wang, B. Zhang, and Y. Zhang. Determinants of public acceptance of tiered electricity price reform in China: Evidence from four urban cities. *Applied Energy*, 91(1):235 – 244, 2012.
 - [56] X. Yang, F. Yu, and W. Pedrycz. Long-term forecasting of time series based on linear fuzzy information granules and fuzzy inference system. *International Journal of Approximate Reasoning*, 81:1 – 27, 2017.
 - [57] J. Zhang, Y.-M. Wei, D. Li, Z. Tan, and J. Zhou. Short term electricity load forecasting using a hybrid model. *Energy*, 158:774 – 781, 2018.
 - [58] L. Zhang and Y. Wu. Market segmentation and willingness to pay for green electricity among urban residents in China: The case of Jiangsu Province. *Energy Policy*, 51:514 – 523, 2012.