Association Rule Mining Based Quantitative Analysis Approach of Household Characteristics Impacts on Residential Electricity Consumption Patterns

Fei Wang[1, 2, *], Kangping Li[1, *], Neven Duić[3], Zengqiang Mi[1], Bri-Mathias Hodge[4], Miadreza Shafie-khah[5], João P. S. Catalão[5, 6, 7]

1. State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (North China Electric Power University), Baoding 071003, China;
2. Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
3. University of Zagreb, Faculty of Mechanical Engineering and Navala Architecture, Ivana Lučića 5, 10000 Zagreb, Croatia
4. National Renewable Energy Laboratory, Golden, CO 80401 USA
5. C-MAST, University of Beira Interior, Covilhã 6201-001, Portugal
6. INESC TEC and the Faculty of Engineering of the University of Porto, Porto 4200-465, Portugal
7. INESC-ID, Instituto Superior Técnico, University of Lisbon, Lisbon 1049-001, Portugal

***Abstract*—The comprehensive understanding of the residential electricity consumption patterns (ECPs) and how they relate to household characteristics can contribute to energy efficiency improvement and electricity consumption reduction in the residential sector. After recognizing the limitations of current studies (i.e. unreasonable typical ECP (TECP) extraction method and the problem of multicollinearity and interpretability for regression and machine learning models), this paper proposes an association rule mining based quantitative analysis approach of household characteristics impact on residential ECPs trying to address them together. First, an adaptive density-based spatial clustering of applications with noise (DBSCAN) algorithm is utilized to create seasonal TECP of each individual customer only for weekdays. K-means is then adopted to group all the TECPs into several clusters. An enhanced Apriori algorithm is proposed to reveal the relationships between TECPs and thirty five factors covering four categories of household characteristics including dwelling characteristics, socio-demographic, appliances and heating and attitudes towards energy. Results of the case study using 3326 records containing smart metering data and survey information in Ireland suggest that socio-demographic and cooking related factors such as employment status, occupants and whether cook by electricity have strong significant associations with TECPs, while attitudes related factors almost have no effect on TECPs. The results also indicate that those households with more than one person are more likely to change ECP across seasons. The proposed approach and the findings of this study can help to support decisions about how to reduce electricity consumption and CO_2 emissions.**

***Keywords*—Electricity consumption pattern; Household characteristics; Association rule mining; Clustering; Apriori algorithm**

Nomenclature

Abbreviations		*Parameters*	
ECP	Electricity consumption pattern	$p(t)$	Actual active power at time t
TECP	Typical electricity consumption pattern	$p^*(t)$	Normalized active power at time t
DBSCAN	Density-based spatial clustering of applications with noise	ε	Radius
HC	Household characteristic	*MinPts*	Minimum number of points
ARM	Association rule mining	$\varepsilon_{initial}$	The initial value of ε
CIE	Chief income earner	$\Delta\varepsilon$	Iterative step
POESL	Proportion of energy-saving lights	χ^2	Test statistic for Chi-squared test
PODGW	Proportion of double glazed windows	$Sup(A \cup B)$	The support of an association rule
NU	Number	$Conf(A \rightarrow B)$	The confidence of an association rule
LHS	Left hand side	$Lift(A \rightarrow B)$	The lift of an association rule
RHS	Right hand side	$Imp(A \rightarrow B)$	The improvement of an association rule
CC	Contingency Coefficient		

1. Introduction

1.1. Background and motivation

Electricity has become an increasingly important energy source for the residential sector in the past few decades. It is estimated by International Energy Agency (IEA) that the share of total electricity consumption in this sector in Organization for Economic Co-operation and Development (OECD) countries has increased from approximately 24.2% in 1974 to 31.1% in 2015 [1]. Although the energy efficiency of home appliances has been significantly improved in recent years, the average electricity consumption of household end-uses in European Union-27 Countries (EU-27) still increased by about 2.5% per year in this period [2]. Therefore, more effective and targeted measures are needed to achieve the EU 20-20-20 energy goals for energy efficiency improvement and CO_2 emissions reduction [3], which requires the comprehensive understanding of residential electricity consumption patterns (ECPs) and how they relate to household characteristics (HCs). The HCs in this paper mainly refer to the characteristics of dwelling, home appliances, occupants and their behaviors. How to identify the most significant HCs affecting the residential ECPs and reveal the complex relationship between them have become the essential problems to support decisions about how to reduce electricity consumption and CO_2 emissions.

Fortunately, regarded as a basic step forward to smart grid, the smart meter installations have increased worldwide in recent years [4]. For example, about 2.2 million smart meters will be installed across the country in Ireland by the end of 2020 [5]. The large-scale deployment of smart meters has enabled the accumulation and storage of electricity consumption data, which provides prerequisites for the study of understanding residential ECP and how they relate to HCs. The knowledge derived from the study can not only help to improve energy efficiency[6], but also contribute to improving tariff design [7], load forecasting and distribution network planning[8-10], and demand side management strategies [11-13].

1.2. Literature review

Clustering has been the most common technique to characterize the behaviors of electricity customers and find representative ECPs in the literature. Various algorithms have been utilized to perform ECP clustering, such as K-means, K-mediods, Fuzzy C-means, hierarchical clustering, follow the leader, ant colony clustering, self-organizing maps (SOM) and Dirichlet process mixture model [14-22]. Actually, in addition to clustering algorithm, typical electricity consumption pattern (TECP) extraction is also important for ECP clustering. As the input objects processed by the clustering algorithm, the TECP of each customer should be created before clustering. Variant TECPs extracted via different methods will inevitably lead to varied ECP clustering results. The most common method obtaining the TECP of individual customer in current studies is to calculate the average value of all the load profiles within a specific period (e.g. a month or a season) [14-16].

Studies using the combination of electricity consumption data and survey information aiming to analyze the relationships between ECPs and HCs are increasing in recent years. Rhodes et al. [23] obtained two different ECPs from 103 households via clustering first and conducted the correlation analysis between profiles and HCs using the binomial probit regression subsequently. The authors found that factors such as if someone works from home, hours of television watched per week, and education levels have significant correlations with average profile shape. This work fills the knowledge gap by identifying correlations between electric customer survey data and electricity use profiles. However, the only two clusters results drawn from a relatively small size of dataset implies that it needs to be further validated for statistically significance with some other large scale data resources. McLoughlin et al. [24] presented a clustering methodology for creating a series of representative electricity load profile classes and linked them with HCs using multi-nominal logistic regression. Beckel et al. [25] estimated the characteristics of household based on supervised machine-learning techniques using electricity consumption data. Viegas et al. [26] also proposed a machine learning based methodology for the classification of new electricity customers and discovery of the drivers of different electricity consumption profiles.

1.3. Research limitations

Understanding the specific influences of various HCs on residential ECPs is challenging because both ECP clustering and association analysis need to be considered. As the literature review shows, even though there are many studies on ECP clustering and a few other works making preliminary efforts on the association analysis between HCs and residential ECPs, some limitations of these studies can be found.

In terms of ECP clustering, the method of how to form the reasonable TECP of each individual customer in a specific period is one of the crucial steps. The TECP of each customer in a specific period should be the most representative ECP in that period,

which can truly reflect customer's typical electricity consumption behavior. However, the current TECP extraction method i.e. average method usually mixes many dissimilar patterns of electricity use together and leads to an unreal reflection of how electricity is actually consumed in reality.

In terms of association relation analysis method, regression and machine learning are the two most common methods. Multicollinearity will occur in most regression models if two or more predictors are highly correlated [27]. Multicollinearity means partial coefficients vary remarkably (sometimes perhaps even change from positive to negative or conversely) while small changes occurred in predictors or datasets, which makes regression models unreliable. Regarding the machine learning methods, the results obtained by these methods can indicate how well the entire bundle of predictors predicts the response variable but are unable to provide detailed information about the cause-effect relationships between explanatory variables and explained response variable.

In terms of the completeness of study, many HCs that potentially have impacts on ECPs have not been investigated in the existing studies. For example, attitudes towards electricity consumption related factors are not included in the existing literature. Whether these factors have impacts on ECPs is still unclear. Additionally, what are the key HCs driving different ECPs is still ambiguous and the explanations of how they work mechanistically are not complete.

1.4. Contributions and paper structure

Facing the above issues, an association rule mining (ARM) based quantitative analysis approach of HCs impact on residential ECPs is proposed to address them together in this paper. The main contributions of this paper can be summarized as follows:

(1) An adaptive density-based spatial clustering of applications with noise (DBSCAN) based TECP extraction method considering the notable discrepancy of the same customer's load profiles with respect to different time periods is proposed in this paper.

(2) An enhanced Apriori based ARM method without the issues including multicollinearity and low explanatory ability existing in those regressions and machine learning models is proposed to reveal the relations between HCs and ECPs.

(3) A massive set of data with various types of HCs is investigated to provide a systematical analysis of the impact of HCs on ECPs in more comprehensive and in-depth perspectives.

The proposed approach and the findings of this research are useful for multi-stakeholder including customers, utilities, and policy makers. Customers can benefit from more targeted, customized and comprehensive energy services. The knowledge obtained by the proposed approach can help utilities to find suitable customer groups for specific energy efficiency programs, design tailor-made tariff schemes, offer directed electricity savings advice, improve the forecasting accuracy and estimate the ECPs of new customers. Policy makers can benefit from the insight into customers' electricity consumption habits to support policy-making for effective energy reduction. More detailed applications of this study will be discussed in Section 5.2.

The paper is organized as follows. Section 2 is the description of the dataset used in this paper. The methodology including three steps is illustrated in section 3. In section 4, the simulation results of TECP extraction, ECP clustering and ARM are presented and analyzed. Section 5 discusses the findings and potential applications of this study. Section 6 highlights the concluding remarks and further works in future.

2. Description of data set

The data used in this research is obtained from the Commission for Energy Regulation (CER) in Ireland [28]. CER carried out the Smart Metering Electricity Customer Behavior Trials (SMECBTs) during 2009 and 2010 for the purpose of assessing the impact on consumer's electricity consumption. Over 4,000 Irish residential customers participated the trials with an electricity smart meter installed in their homes and agreed to finish a comprehensive survey concerning electricity consumption behaviors, such as home sizes, lifestyles and altitudes towards energy saving.

2.1. Smart metering dataset

The smart metering dataset is comprised of electricity consumption data of 4232 residential customers at 30 minutes interval over one and a half year. No personal or confidential information is contained in the dataset, which instead gives the behavioral and usage patterns anonymously. We select the data of a full year from January 1st to December 31st, 2010 and the dataset is reduced to 3644 customers after removing 588 customers with spurious and missing data.

2.2. Surveys dataset

The surveys consist of 143 questions containing information including the characteristics of the dwellings (e.g. dwelling type, year of construction), the socio-demographic data (e.g., age of householder, number of occupants), the appliance's ownership and heating (e.g. number of appliances and house heating) and the attitudes towards energy (e.g. the willing to reduce energy use). Every record of each customer in the surveys dataset is linked to the corresponding records in the smart meter dataset through the unique ID of each customer. We are able to identify 3427 valid surveys by the ID among these 3644 customers. Hence, the sample size is finally trimmed to 3427 customers for the further analysis in this paper.

2.3. Overview and classification of HCs

All HCs included in the survey are summarized in Table 1 and can be divided into four categories: dwelling characteristics, socio-demographic, appliances and heating and attitudes towards energy.

The numbers with bold fonts in Table 1 denote the proportion of those customers who answered the corresponding questions. The sum of proportion in some rows is less than 100% because some customers refused to answer the questions. Each response to every question is treated as an individual item in the subsequent association rules analysis.

For the dwelling characteristics factors, we note that more than half of customers gave an invalid answer in the question of total floor area. So it is not included in this paper.

Regarding the socio-demographic factors, it is noted that social classes of the chief income earner (CIE) are coded as 'AB', 'C', 'DE' and 'F'. 'AB' represents high and intermediate managerial, administrative or professional occupation. 'C' represents supervisory and clerical, junior managerial and skilled manual workers. 'DE' indicates semi-skilled and unskilled manual workers, state pensioners, or unemployed. 'F' represents farmers. Although approximately half of customers gave a rejection answer about their income, we still select this factor for the analysis because this factor has been reported to be significantly associated with electricity consumption [29].

Table 1 Overview of HCs

Categories and Factors	Responses (proportion/%)
dwelling characteristics	
Dwelling type	Apartment(**1.65**)|semi-detached(**29.68**)|detached(**27.51**)| terraced(**13.92**)|bungalow(**27.03**)
Dwelling age	<10years(**20.84**)|10~30(**28.68**)|30~75(**38.21**)|>75years(**12.27**)
No. of bedrooms	1~2(**9.11**)|3(**43.60**)|4(**35.48**)|5+(**11.55**)}
POESL[a]	None(**21.26**)|quarter(**26.49**)|half(**17.05**)|three quarters(**16.93**)|all(**18.28**)
PODGW[b]	None(**8.00**)|quarter(**1.98**)|half(**2.98**)|three quarters(**2.62**)|all(**84.43**)
Insulated walls	Yes(**57.40**)|no(**31.30**)|don't know(**11.30**)
socio-demographic	
Sex of respondent	Male(**50.69**)|female(**49.31**)
Age of respondent	18~35(**9.05**)|36~55(**44.35**)|>55(**45.94**)
Employment status[c]	An employee(**46.00**)|self-employed(**12.54**)|unemployed(**9.47**)|retired(**31.99**)
Social class[c]	AB(**14.67**)|C(**43.17**)|DE(**38.51**)|F(**2.56**)
Occupants	live alone(**19.36**)|all occupants are adults(**53.04**)|both adults and children(**27.60**)
Education level[c]	No formal education(**1.38**)|Primary(**11.49**)|secondary level(**45.25**)|third level(**36.38**)
Income[c]	<30k(**29.04**)|30k~75k (**25.5**)|>75k(**8.51**)|refused(**36.95**)
appliances and heating	
NU[d]. of washing machines	0(**1.68**)|1(**97.59**)|2(**0.73**)
NU. of tumble dryers	0(**31.39**)|1(**68.46**)|2(**0.15**)
NU. of dishwashers	0(**33.13**)|1(**66.63**)|2(**0.24**)
NU. of electric showers	0(**31.27**)|1(**62.90**)|2+(**5.83**)
NU. of electric cookers	0(**23.39**)|1(**76.25**)|2+(**0.36**)
NU. of electric heaters	0(**68.73**)|1(**24.29**)|2(**6.98**)
NU. of stand alone freezers	0(**49.19**)|1(**48.86**)|2+(**1.95**)

NU. of water pumps	0(**80.71**)|1(**18.82**)|2(**0.47**)
NU. of immersions	0(**23.37**)|1(**76.28**)|2(**0.35**)
NU. of TV<21inch	0(**34.91**)|1(**39.39**)|2+(**25.71**)
NU. of TV>21inch	0(**15.63**)|1(**50.72**)|2+(**33.64**)
NU. of desktop computers	0(**51.98**)|1(**45.01**)|2+(**3.01**)
NU. of laptops	0(**46.51**)|1(**41.91**)|2+(**11.58**)
NU. of game consoles	0(**66.90**)|1(**22.16**)|2+(**10.94**)
Cook by electricity	Yes(**69.51**)|no(**30.49**)
House heating by electricity	Yes(**7.16**)|no(**92.84)**
Water heating by electricity	Yes(**57.48**)|no(**43.27**)
attitudes towards energy	
Be interested in changing electricity use if it reduces the bill	1[c](**84.10**),2(**10.91**),3(**3.03**),4(**0.96**),5(**0.99**)
Be interested in changing electricity use if it helps the environment	1(**75.98**),2(**16.92**),3(**4.64**),4(**1.43**),5(**1.02**)
It is too inconvenient to reduce our usage of electricity	1(**5.89**),2(**11.76**),3(**12.11**),4(**23.64**),5(**43.95**)
I do not want to be told how much electricity I can use	1(**18.15**),2(**11.91**),3(**14.04**),4(**19.70**),5(**33.56**)
I have already done a lot to reduce electricity use	1(**34.26**),2(**31.78**),3(**19.11**),4(**10.10**),5(**4.76**)

a. POESL: Proportion of energy-saving lights
b. PODGW: Proportion of double glazed windows
c. All of these questions are designed for the chief earner of the household
d. NU: Number
e. The answer scale: 1-strongly agree; 2-tend to agree; 3-neither agree nor disagree; 4-tend to disagree; 5-strongly disagree

As for the appliances and heating related factors, all of the appliances can be divided into two general categories: "wet" appliances (i.e. water related appliances, e.g. washing machine, dishwashers and immersions) and entertainment appliances (e.g. TV, computer and game console). Noting that most Irish households do not use electricity for home heating and do not have air conditioning [30].

The attitudes towards energy related factors are not included in most literature on the same topic. Why we consider these factors in our study is that the interventions from different electricity consumption behaviors can lead to distinct ECPs. These factors are divided into three categories: attitudes towards reduction (the first two questions), reasons for not reducing usage (the third and fourth questions) and energy reduction efforts (the last question).

3. Method

The proposed approach shown in Fig. 1 is divided into three steps. First, four TECPs of each individual customer corresponding to four seasons are extracted using only weekday data with an adaptive DBSCAN algorithm. Second, all TECPs extracted from step 1 are separately grouped into several non-overlapping clusters over each season according to the similarity using K-means algorithm. Third, an enhanced Apriori algorithm is proposed to find the association rules between TECPs and thirty-five factors covering four categories of HCs including dwelling characteristics, socio-demographic, appliances and heating and attitudes towards energy, and then the explanation of functioning mechanism is also presented afterwards.

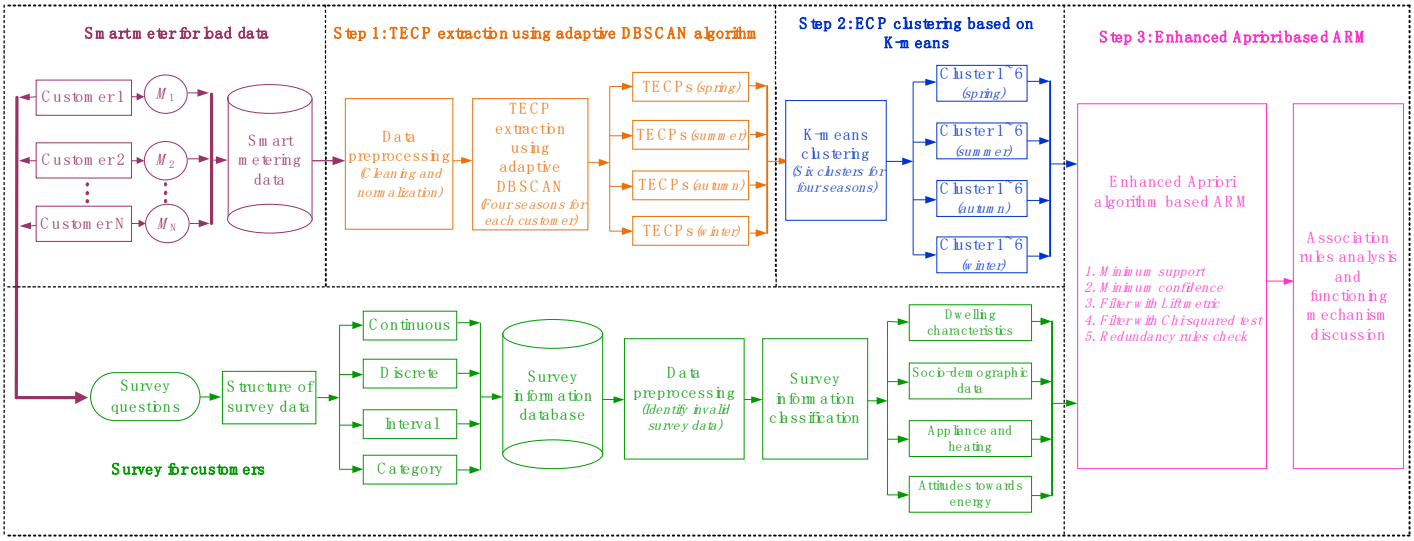


Fig. 1. Framework of the proposed approach based on enhanced Apriori algorithm

3.1. Data normalization and season divison

In order to characterize the patterns of electricity use, the influence caused by different amplitudes in load data needs to be eliminated through the normalization before clustering. Each daily load data is normalized to the corresponding total daily electricity consumption value by formula (1) [18].

$$p^*(t) = \frac{p(t)}{\sum_{t=1}^{48} p(t)} \tag{1}$$

where $p(t)$ and $p^*(t)$ are the actual and normalized active power at time t.

On top of the normalization, the conspicuous discrepancy of residential ECPs between weekdays and weekend days should also be taken into account. Due to the space limitation of this paper, we only use the load data in weekdays because probably it is somehow more typical and important than those in weekend days. Another issue affecting ECPs needed to be considered is the seasonal effects, so the whole year is divided into four seasons: spring, from March 01 to May 31; summer, from June 01 to August 31; autumn, from September 01 to November 31 and winter from December 01 to December 31 combine with January 01 to February 28 [31].

3.2. Extraction of TECP

There are considerable variabilities in the load profiles for a given residential customer throughout a whole year due to the different weather conditions and electricity consumption behaviors. To depict its characteristics, the TECP of each customer needs to be extracted from load data. Considering the trade-off between accuracy and complexity, four TECPs of each customer in corresponding season are defined respectively as the most representative ECP in the corresponding duration.

To extract the TECP of each customer, those ECPs occurring frequently and other uncommon ECPs should be identified and separated. DBSCAN is a well-known density based data clustering algorithm, which can find arbitrarily shaped clusters and also handle outliers effectively [32]. It defines a cluster as an area that has a higher density than its surrounding areas, which is consistent with the definition of the TECP (i.e. common ECPs have higher density than the other uncommon ones). Those uncommon ECPs are considered as *noise* and can be effectively identified by the DBSCAN. Here why we choose DBSCAN instead of other well-known clustering algorithms (e.g. K-means, FCM) is because the DBSCAN can handle outliers effectively while K-means/FCM has the low ability to isolate outliers [14].

DBSCAN includes two parameters: radius ε and minimum number of points *MinPts*. The algorithm checks the ε-neighborhood of each point. A point is marked as a *core point* if the number of points within its radius ε (including itself) is equal to or larger than *MinPts* and then a new cluster is created. Further points are added iteratively to the cluster by finding *core points* for each point in the ε-neighborhood of the cluster. The algorithm terminates if no more points exist that can be assigned to a cluster. Points that are reachable for core points but are not core points are marked as *border points* and also added to the cluster.

Those points that cannot be assigned to any cluster during the algorithm are marked as *noise*. The illustration of these three types of points is shown in Fig. 2 [33].

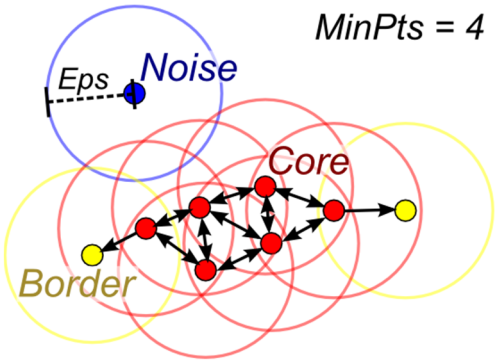


Fig. 2. Illustration of core points, border points and noise points

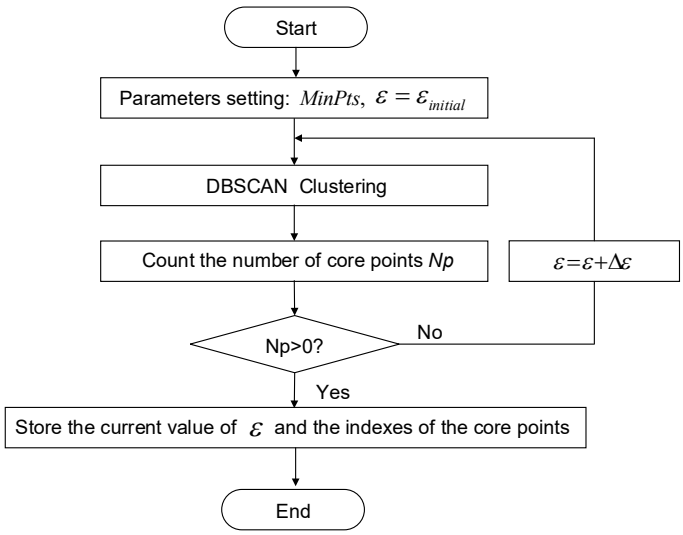


Fig. 3. Flow chart of the proposed adaptive DBSCAN algorithm

One of the disadvantages of DBSCAN is the difficulty in selection of the parameters. Apparently, the clustering result of DBSCAN is highly related to these two parameters. It is unreasonable to use fixed parameters of DBSCAN to extract TECP for all customers because the distribution of ECPs varies for different customers due to different personal lifestyles and consumption behaviors. Therefore, an adaptive DBSCAN is proposed in this paper to dynamically adjust the parameter radius ε to improve the ability to accommodate any possible distribution of load data. The detailed procedure of the proposed adaptive DBSCAN algorithm is shown in Fig. 3. Euclidean distance is chosen as the distance metric for this algorithm. It starts from $\varepsilon = \varepsilon_{initial}$ and automatically adjusts the value of ε with the iterative step $\Delta\varepsilon$ until the number of core points is non-zero.

For each customer, after obtaining a number of load profiles marked as *core points*, a single TECP is then created by averaging those load profiles that marked as core points. Moreover, the final ε value is stored and can be used to characterize the regularity of every customer's electricity consumption behavior. The smaller the ε is, the more regular the consumption behavior is. Those customers with high values of ε are excluded from the dataset. Finally, the total remaining sample size is 3326 after trimming. The details of parameter setting and results of TECP extraction will be presented in Section 4.1.

3.3. Load pattern clustering

ECP clustering refers to segmenting customers into several clusters such that customers in the same cluster show similar ECPs while customers in different clusters exhibit distinct ECPs. K-means, the most widely used clustering algorithms, is adopted for ECP clustering due to its attractive advantages such as fast computation speed and effective clustering results. K-means groups a

set of unlabeled data into K clusters through an iterative process to minimize the sum of square error for all clusters. K is given before clustering. The centroid of each cluster is obtained by calculating the average value of all the data points in the same cluster.

The determination of K and the selection of initial centroids are the two main difficulties of K-means. To find a suitable value of K, two indexes applied in many works [14-16], Davies-Bouldin index (DBI) and Ratio of Within Cluster Sum of Squares to Between Cluster Variation (WCBCR), are adopted to evaluate the validity of clustering. For these two indexes, smaller values indicate better clustering results. However, excessive subdivision clustering is not suitable and helpful for application. Hence, the optimal number of clusters should be determined based on both these indexes and the specific purpose of the clustering. For the second issue, K-means is conducted for a large number of times and the initial cluster centroids are generated randomly in each round. The record with the best clustering indexes among all these cases is chosen as the final result.

3.4. Association rules mining

3.4.1. Standard Apriori algorithm

ARM is widely used to discover interesting relationships among items within a given dataset. Apriori, the most popular ARM algorithm, was first introduced by Agrawal et al. in 1993 for the purpose of finding association relations between different transactions in a large Boolean transactional database [34].

Let $I=\{i_1,i_2,...,i_m\}$ be a set of items. Each transaction T is a set of items, such that $T \subseteq I$. An association rule is of the form $A \rightarrow B$, where the left hand side (LHS) 'A' is a set of items referred to as the antecedent of the rule, and the right hand side (RHS) 'B' is a set of items referred to as the consequent of the rule. This rule indicates that the occurrence of B can be predicted based on the occurrence of A.

The *support* of an association rule denoted by *Sup*($A \cup B$) is the proportion of those transactions containing both A and B in the whole set. The *confidence* of an association rule is defined as the proportion of those transactions including both A and B in the transactions containing A, expressed in formula (3).

$$Conf(A \rightarrow B) = \frac{Sup(A \cup B)}{Sup(A)} \tag{3}$$

Support and *Confidence* are two important indexes to evaluate the interestingness of a rule. The rules satisfying user's threshold (minimal *Support* and minimal *Confidence*) are considered as interesting rules in standard Apriori algorithm. However, the rules set generated by minimum *support* and minimum *confidence* constraints are often too numerous to be utilized efficiently. Many rules are often redundant and unrelated, which will produce adverse effect on results [35]. Thus, an enhanced Apriori algorithm with additional interestingness measures is proposed to solve the above problems of current standard Apriori algorithm probably causing misleading rules, redundant information, random and coincidentally occurring rules.

3.4.2. Enhanced Apriori algorithm

Let $\boldsymbol{HC}=\{hc_1,hc_2,...hc_m\}$ denotes the HCs set presented in section 2.3. It is noted that not all the factors in $\boldsymbol{HC}$ are significantly associated with customers' ECPs. Thus, a HCs preliminary selection procedure is carried out before performing Apriori. The Chi-squared test of independence can be used to determine if there is a significant relationship between two categorical variables [36]. This test utilizes a contingency table to analyze the data. For example, the contingency table of variables 'age of respondent' and 'clustering result for spring' is shown in Table 2.

Table 2 The contingency table of variables 'age of respondent' and 'clustering result for spring'

		Clustering result for spring						Total
		cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	
The values of IF "*Age of respondent*"	1	341	324	364	286	126	89	1530
	2	109	54	95	79	50	30	417
	3	93	19	103	42	50	8	315
	4	326	60	260	186	211	21	1064
Total		869	457	822	593	437	148	3326

Let χ^2 denotes the test statistic for the Chi-squared test of independence, which can be computed as formula (4).

$$\chi^2 = \sum_{i=1}^{Rows}\sum_{j=1}^{Cols}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \tag{4}$$

where o_{ij} is the observed cell count in the i^{th} row and j^{th} column of the table. *Rows* is the number of rows and *Cols* is the number of columns. e_{ij} is the expected cell count in the i^{th} row and j^{th} column of the table, which can be calculated as formula (5).

$$e_{ij} = \frac{(\sum_{k=1}^{Cols} o_{ik})(\sum_{k=1}^{Rows} o_{kj})}{N} \tag{5}$$

where N is the total number of customers. The calculated χ^2 value is then compared to the critical value from the χ^2 distribution table with degrees of freedom *df*= (*Rows*-1)(*Cols*-1) and chosen confidence level. If the calculated χ^2 value>critical χ^2 value, there is a significant association between the tested two variables. We used Chi-squared test of independence to test whether there is an association between each factor $hc_j, (j = 1,2,...,m)$ in ***HC*** and clustering result for each season. A 95% confidence level is used and those HCs which have no association with clustering result are removed.

The remaining HCs are ranked by the Contingency Coefficient (*CC*), expressed in formula (6).

$$CC = \sqrt{\frac{\chi^2}{N+\chi^2}} \tag{6}$$

CC can be used to quantify the correlation degree of two categorical variables [37]. Higher values of *CC* indicate higher degrees of correlation.

Apriori algorithm is performed after HCs preliminary selection. The rules set generated by Apriori algorithm is denoted by $\mathbf{R}$. Since we focus on identifying the key HCs of different ECPs, only the rules containing ECP variables as RHS consequent are selected for further analysis and these rules form a new rule subset denoted by $\mathbf{R}'$.

Lift measure is proposed to overcome the disadvantage of *confidence* measure by not taking the baseline frequency of the consequent into consideration, given by formula (7).

$$Lift(A \to B) = \frac{Sup(A \cup B)}{Sup(A)Sup(B)} \tag{7}$$

The *Lift* measure is defined over $[0,+\infty)$. If $Lift = 1$, the occurrence of A is independent of the occurrence of B. Rules with $Lift \le 1$ are removed to ensure the antecedent of the rule has a positive promoting effect on the consequent of the rule. Then a new rules subset named $\mathbf{R}''$ satisfying the above constraints can be obtained.

In order to ensure that the rules extracted are not purely caused by random coincidence, we still use the Chi-squared test (χ^2) to determine if the correlation between the antecedent and consequent of an association rule $A \to B$ is statistically significant. As such, the new rules set $\mathbf{R}'''$ can be obtained after removing the rules with no statistical significance.

Some rules in $\mathbf{R}'''$ with the same consequent but different antecedents probably imply nearly the same knowledge, thus the redundancy should be checked for each rule. Let us denote the *improvement* of a rule $A \to B$ as the minimum difference between its confidence and the confidence of any proper sub-rule with the same consequent [38], given by formula (8).

$$Imp(A \to B) = \min(\forall A' \subset A, Conf(A \to B) - Conf(A' \to B)) \tag{8}$$

where A' represents the sub-set of A. Large positive value of *improvement* indicates that every item or its combination in the antecedent of the rule plays an important role in the predictive ability of the rule. In this paper, we remove those rules whose *improvement* is less than 5% of its own *confidence* to obtain more refined and significant rules set. The flow chart of the proposed enhanced Apriori algorithm is shown in Fig. 4.

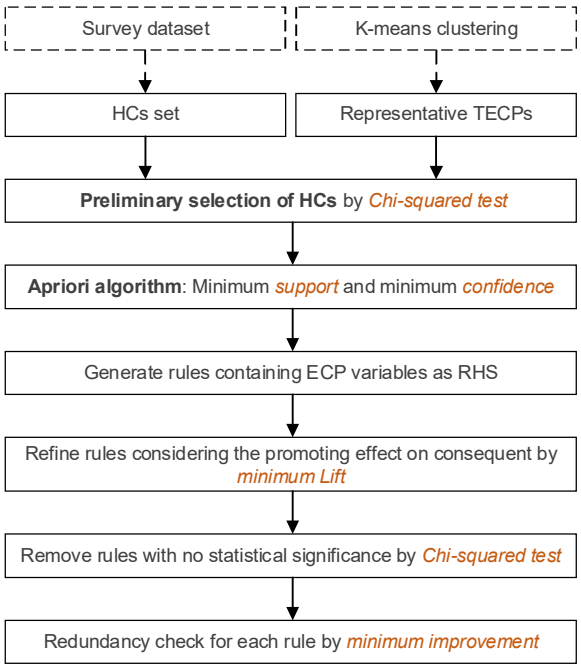


Fig. 4. Flow chart of the enhanced Apriori for ARM.

4. Case study

4.1. Results of TECP extraction

To extract the seasonal TECP of each customer in weekdays, the adaptive DBSCAN algorithm was performed for each season separately. Here we set $\varepsilon_{initial}$ and $\Delta\varepsilon$ to be 0 and 0.02 respectively. The $\Delta\varepsilon$ was set to be small enough to achieve a precise search. The parameter *MinPts* was set according to the number of points for clustering. There are total 66 weekdays in a season, thus we set *MinPts* as 20 (approximately 1/3 of the number of weekdays in a season) in this paper. As such, for each customer n ($n = 1,2,...,N$), four different values of ε ($\varepsilon_n^{spring}, \varepsilon_n^{summer}, \varepsilon_n^{autumn}, \varepsilon_n^{winter}$) corresponding to four seasons can be obtained. The mean value of these four ε was calculated for each customer and denoted by $\overline{\varepsilon_n}$ ($n = 1,2,...,N$).

As an example, the extraction process of the TECP in spring weekdays for customer #1 and #2 are shown in Fig. 5. For customer#1, $\overline{\varepsilon_1}$ is 0.05. The common ECPs of this household appear in 33 days, which accounts for 50% of the spring weekdays. However, for customer #2, $\overline{\varepsilon_2}$ is 0.33, more than six times larger than $\overline{\varepsilon_1}$. It can be clearly observed in Fig. 5 (b) and (e) that the load profiles marked as core points of customer#1 are much more compact than customer#2. In other words, it is hard to identify the TECP of customer#2 because its load profiles are still dispersive even after distinguishing different seasons and weekdays vs. weekend days. The comparisons between the proposed adaptive DBSCAN method and the average method are shown in Fig. 5 (c) and (f). The TECP of customer#1 derived by average method shows a peak demand at about 15:00, however, this peak is caused by combining those uncommon ECPs (i.e. the ECP with high normalized consumption in Fig. 5 (a)) with actual TECP together, which is an unreal reflection of typical electricity consumption behavior for customer#1.

In order to make the clustering results more reliable, those customers with high values of $\overline{\varepsilon}$ were removed from the data set, because it is difficult to extract the TECP for them. The mean value of $\overline{\varepsilon}$ for all customers was 0.132 and the standard deviation was 0.035. Considering the trade-off between accuracy and sample numbers, customers with $\overline{\varepsilon}$ values that were two standard deviations above mean value were excluded from the dataset. The total remaining sample size was $N = 3326$ after removing 101 customers which only accounts for 3% of the population and will not affect the validity of this work. The remaining samples formed the basis for all the analysis carried out in the following sections.

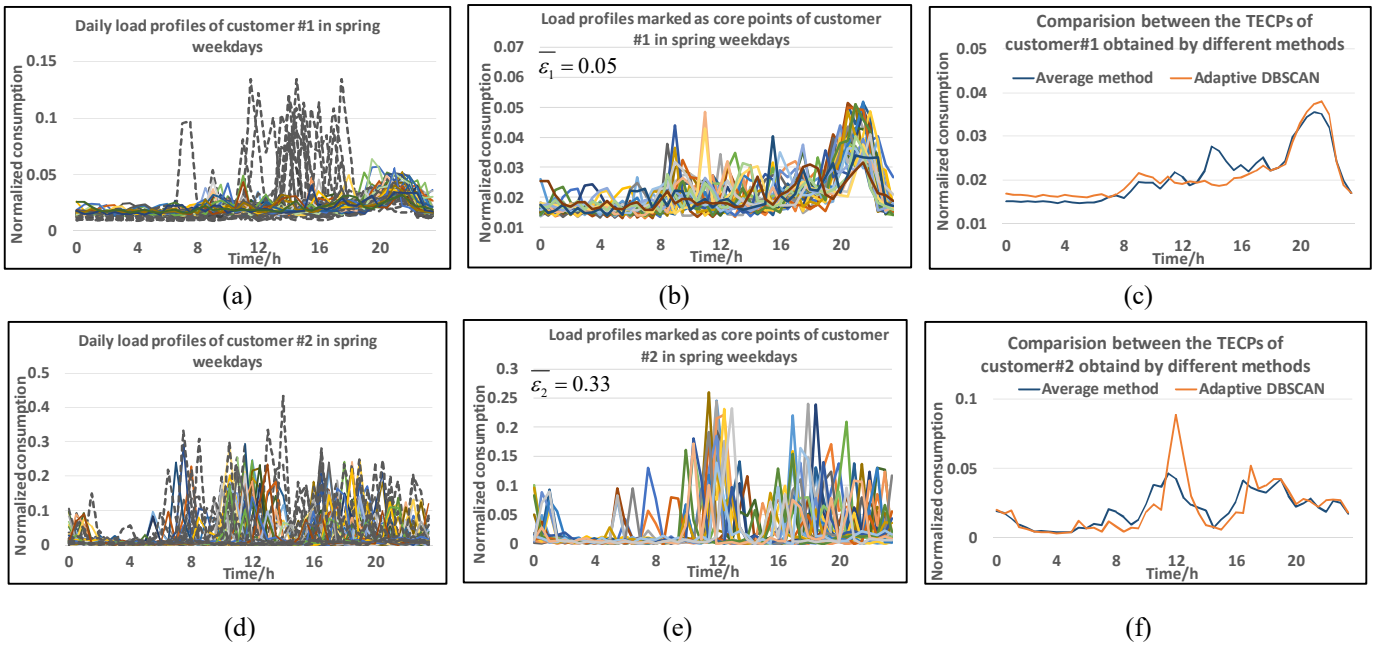


Fig. 5. TECPs of customer#1 and customer#2 for weekdays in spring.

4.2. Results of ECP clustering

K-means was implemented through MATLAB R2012b. Over 50 rounds with the number of clusters ranging from 2 to 12 were performed. A new set of initial centroids was chosen randomly at each round. DBI and WCBCR were calculated for each round, and their average values are shown in Fig. 6.

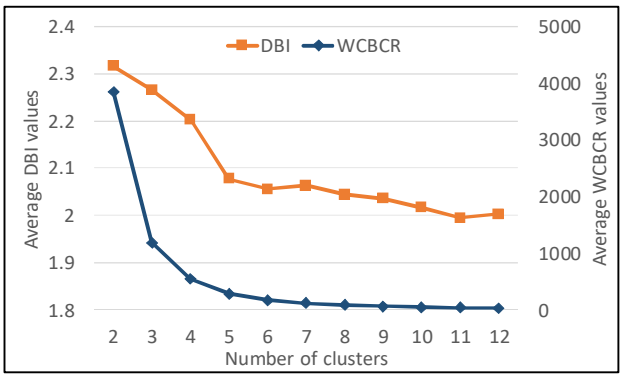


Fig. 6. Two clustering evaluation indexes for k-means for spring.

In general, these two indexes present decreasing tendency as the number of clusters increases. We note that both DBI and WCBCR show a convergence trend when the number of clusters is larger than 5. Two knee points can be observed at 6 and 11 for DBI. So, the optimal number of cluster can be chosen as 6 or 11. Considering the balance between clustering quality and complexity, we finally set the number of clusters to be 6. The same process was performed for the other three seasons as well.

The detailed results of TECP clustering for weekdays in spring are shown in Fig. 7. It can be seen that the six TECP clusters differ from each other in terms of the number of peaks and their occurrence time. In the majority of clusters (e.g. cluster 2, cluster 4, cluster 5 and cluster 6), there are two obvious peaks in the morning and evening. Cluster 3 shows three peaks in the morning, dusk and evening respectively. Cluster 1 presents a flat pattern across the whole day in 24 hours period, which is significantly different from other clusters.

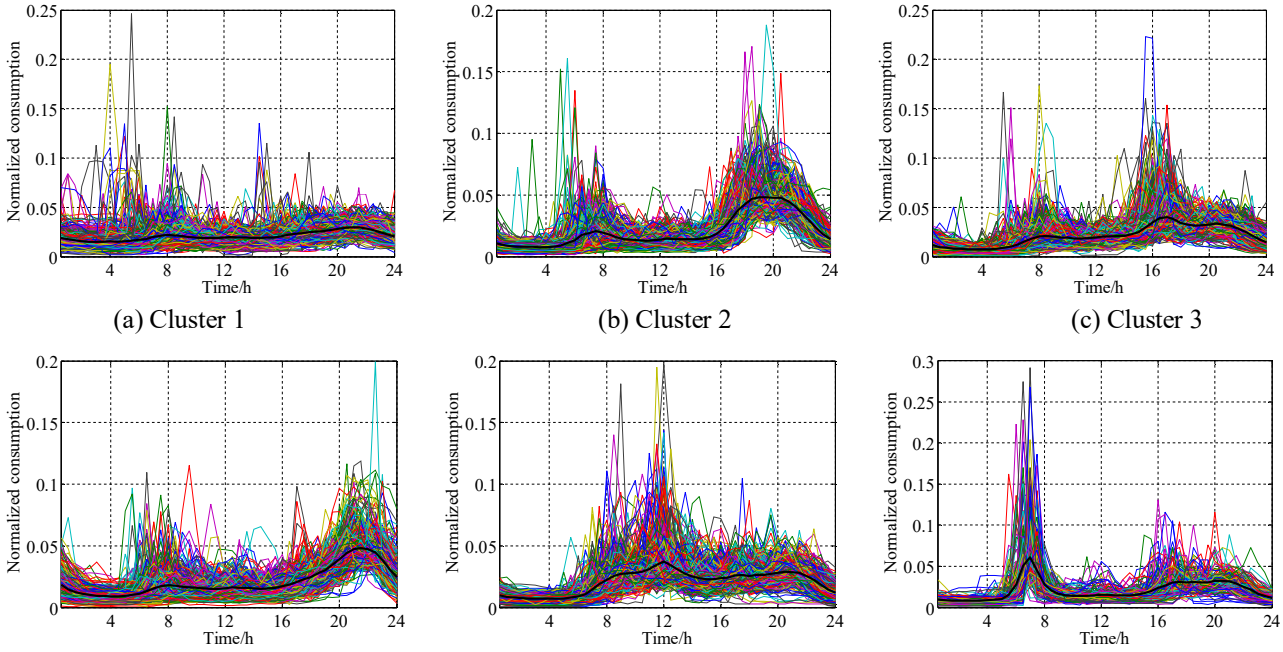


Fig. 7. Clustering results of ECPs for weekdays in spring. The black line in each cluster represents the centroid of that cluster.

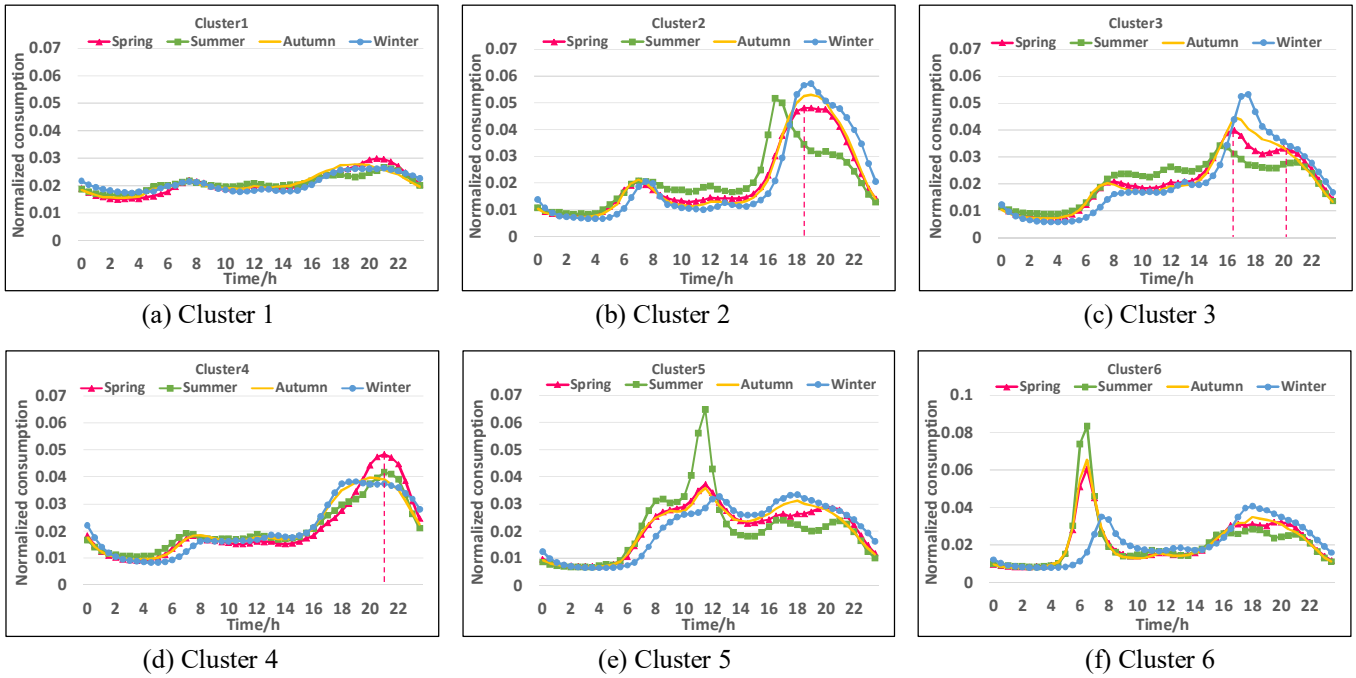
The representative ECPs for all seasons obtained by averaging the TECPs of all customers in the same cluster are shown in Fig. 8. The percentage of number of customers in six clusters for four seasons are summarized in Table 3.

Table 3 The percentage of number of customers in six clusters for four seasons

Cluster	Spring	Summer	Autumn	Winter
1	26.13%	28.47%	19.03%	14.02%
2	13.74%	14.73%	13.98%	11.06%
3	24.71%	27.48%	24.26%	17.38%
4	17.83%	20.93%	25.35%	24.95%
5	13.14%	4.93%	13.53%	17.98%
6	4.45%	3.46%	3.85%	14.61%

Seasons have important impacts on electricity consumption behaviors because of the variations of daytime and weather conditions (e.g. temperature). From Fig. 8 we find that the degree of inter-seasonal effect on ECPs varies from cluster to cluster. For example, cluster 1 shows less variance probably because there are no seasonal appliances and no changes in occupancy for these households. In contrast, other clusters show lager variances in ECPs across seasons. Compared to summer, cluster 4 and 6 present a later demand peak in the morning and an earlier demand peak in the evening in winter. It is most likely related to the change of daytime length across seasons.

Meanwhile, we note that cluster 5 and 6 have a greater difference between evening and morning peak in summer, which is probably caused by the higher temperature difference between day and evening. On the contrary, cluster 2 and 3 show an earlier electricity use peak in the evening for summer, which is probably due to the change of occupancy (e.g. children stay at home during summer holiday).



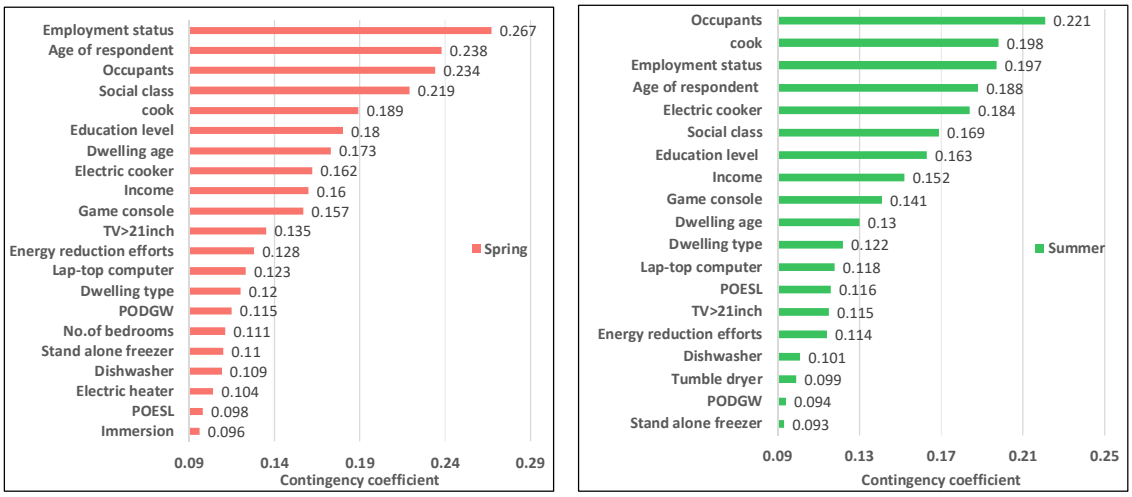
(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

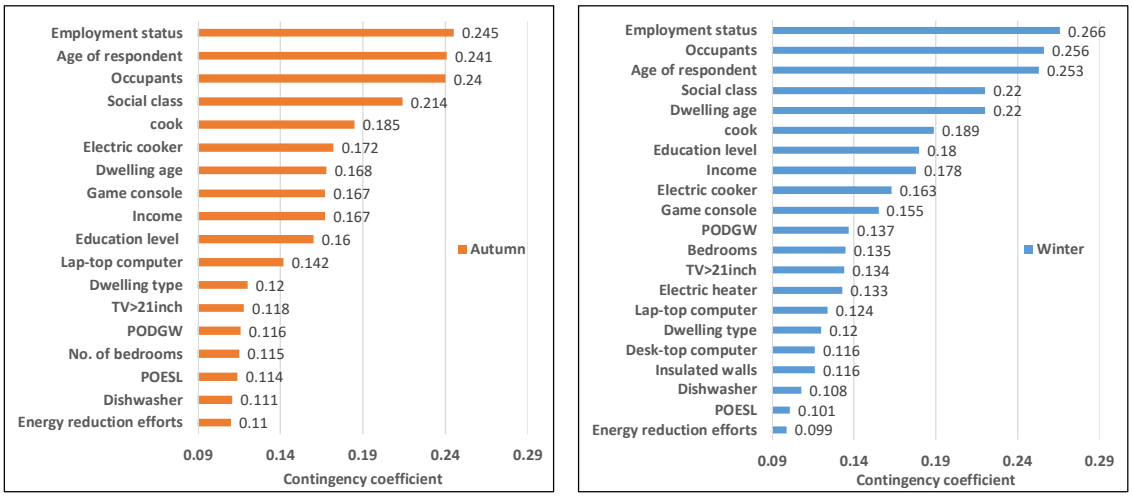
Fig. 8. Six representative ECPs for four seasons. Here each representative ECP represents the centroid of each cluster for each season.

4.3. Results of HCs preliminary selection

As discussed in section 3.4.2, preliminary selection of HCs was performed using a confidence threshold of 95% before ARM. The HCs showing correlations with ECPs but without statistical significance were removed directly from the HCs set and the remaining HCs were ranked by *CC*. The selection and ranking results for four seasons are shown in Fig. 9.



(a) Spring (b) Summer



(c) Autumn (d) Winter

Fig. 9. Results of HCs selection for four seasons

Although the ranking results vary in seasons, we can still find that all of the socio-demographic factors such as *Employment status, Occupants, age of respondent, social class, education level and income*, except for *sex*, are ranked at the top of the *CC* list for all seasons. *Employment status* shows the strongest correlation with ECPs in spring, autumn and winter. Attitudes towards energy related factors present no significant correlations with ECPs, except for *energy reduction efforts*. Regarding the appliances and heating related factors, *cook* shows a significant correlation with ECPs. Hence, it is not surprising that *Electric cooker* also presents a significant correlation in all of the appliances for four seasons. In addition, entertainment appliances such as *Game console, Lap-top and TV>21 inch* are significantly associated with ECPs, whereas no significant relationship is found between ECPs and water related appliances such as *washing machine, water pump and Immersion.* This finding is inconsistent with the conclusion in the literature [26]. Interestingly, *Insulated wall* shows a significant effect on ECPs in winter, but no effect in other seasons, which is probably due to the low temperature in winter.

4.4. Association rules analysis

For each customer, the remaining HCs after selection and the corresponding ECP formed an item set, namely *item set0*. The combination set compose of all these items is the process objects of ARM. The Enhanced Apriori algorithm is applied to find if there are any significant correlations between HCs and certain ECPs of customers. Relative minimum support has been used to avoid the problem that no rule associated with a specific ECP due to the imbalance distribution of clusters. Namely, the minimum support was set according to the proportion of each cluster in population. Considering the number of the rules generated and the reliability of the rules, the relative minimum support was finally set to be 0.3 in this paper. The minimum confidence was set to be the proportion of each cluster in population so that the rules satisfying the minimum confidence constrain also satisfy the *Lift* constraint. We used the statistical program R (version 3.3.2) to implement the enhanced Apriori algorithm. The algorithm was performed separately for each cluster and each season (i.e. totally 24 executions) on a standard PC with an Intel ® Core ™ i7-5500U CPU @ 2.40 GHz, and 8.0 GB RAM. The average running time of each execution is about 2.46 seconds. A 95% confidence level was applied to these executions to pick out the rules with statistically significance. The rules corresponding to each cluster for spring are sorted by *Lift* values and summarized in Table 4. Taking the spring as an example, the rules obtained are illustrated as follows.

4.4.1. ECP cluster 1 (ECPC1)

In Fig. 7, ECPC1 describes a flat usage with no distinct peak across the 24h period. In Table 4, there are 8 rules for ECPC1 after refining the rules not satisfying constraints. Rule 1, 3, 5, 6 and 7 indicate the significant association between socio-demographic related factors and ECPC1. Rule 1 describes that occupiers living alone have the strongest association with this cluster. Additionally, households with a CIE of the third education level, or with social class of "DE" and have no Game console are more likely to use electricity like this cluster as presented in rule 3 and 5. As illustrated in rule 6 and 7, older households

(>55years) and retired people are more likely to belong to ECPC1 compared with the other people. Rule 2 implies that those households cooking not by electricity show significant associations with ECPC1 and it is not surprising that households without electric cooker are also more likely to belong to this pattern. Finally, as stated in rule 8, households without dishwasher are also more likely to belong to this cluster.

4.4.2. ECP cluster 2 (ECPC2)

ECPC2 reflects two distinct electricity consumption peaks occurring in the morning and night, respectively. There are 53 rules related to this cluster, but only 10 rules with top 10 largest *Lift* values are presented. Among all these rules, "Employment=an employee" appears 8 times, which strongly indicates that householders working as an employee are more likely to belong to ECPC2. Other socio-demographic factors such as middle aged households (age of respondent=36~55 years), third level education, social class of "C", living with both adults and children, dwelling with 4 bedrooms also show significant associations with this cluster. Regarding the appliance related factors, the households cooking by electricity and owning a dishwasher but without any stand alone freezer, Game console or electric heater are more likely to consume electricity like ECPC2.

4.4.3. ECP cluster 3 (ECPC3)

ECPC3 shows three peaks in the morning, dusk and evening, and is more obvious for spring and summer. Compared to ECPC2, the first evening peak appears earlier at about 5:00 pm and gradually declines until the next peak appears at about 8:00 pm. Similarly, we also selected the 10 rules with the largest *Lift* value from the 28 rules. "Cook by electricity" appears 6 times in the 10 rules. As indicated in rule 1, the households with a CIE of the secondary education level, owning a dishwasher and cooking by electricity show the strongest association with this cluster. In contrast to ECPC2, the households owning an electric heater, stand alone freezer or two or more TVs which are bigger than 21 inch are more likely to belong to this cluster. Besides, people living with all adults in a 30~75 years old dwelling and with 3 bedrooms show significant associations with ECPC3.

4.4.4. ECP cluster 4 (ECPC4)

Similar to ECPC2, ECPC4 also shows two peaks but the evening one is later than ECPC2 at about 9:30 pm. Only 5 rules related to this cluster can be found. Cooking not by electricity exhibits the strongest association with this cluster as stated in rule1, thus not surprisingly the households having no electric cooker also present significant associations with ECPC4. People living alone in a bungalow house tend to use electricity like ECPC4. The households that do not own a freezer but have a TV>21 inch are more likely to belong to this cluster.

4.4.5. ECP cluster 5 (ECPC5)

ECPC5 presents a quiet high peak around midday and relatively smaller peak in the evening. Similar to ECPC1, retired households (age of respondent>55) with social class of "DE" show strong associations with this cluster. In contrast to ECPC1, people that live with all adults and cook by electricity tend to use electricity like this cluster. In particular, the households with yearly income of CIE less than 30k euros are more likely to belong to ECPC5. Regarding the appliances type, the households that do not own laptop computers and dishwashers but have an electric cooker show significant associations with this cluster.

4.4.6. ECP cluster 6 (ECPC6)

ECPC6 showing a large morning demand peak and relatively smaller demand peak in evening is quite different from other clusters. The number of households in this cluster is less than any other cluster. The households with a CIE of the third education level and owning a TV>21 inch present the strongest significant associations with this cluster as implied in rule1. Similar to ECPC2, middle aged households with a CIE working as employees and live with both adults and children are more likely to use electricity like ECPC6. Besides, the households owning electric heaters tend to belong to this cluster.

Table 4 Summary of the rules obtained by enhanced Apriori algorithm for spring season

Rules	LHS	RHS	Sup(%)	Conf(%)	Lift
	Rules related to ECPC1				
1	{Occupants=live alone}	{ECPC1}**	7.49	38.66	1.48
2	{Cook=no}	{ECPC1}**	10.17	34.98	1.34

3	{Education=third level &Game console=0}	{ECPC1}**	7.33	32.82	1.26
4	{Electric cooker=0}	{ECPC1}**	7.60	32.65	1.25
5	{Social class=DE& Game console=0}	{ECPC1}**	9.56	30.93	1.18
6	{Employment=retired}	{ECPC1}**	9.80	30.64	1.17
7	{Age of respondent=55+}	{ECPC1}**	13.65	29.71	1.14
8	{Dishwasher=0}	{ECPC1}*	9.38	28.31	1.08
	Rules related to ECPC2				
1	{Employment=An employee & Dishwasher=1&freezer=0}	{ECPC2}**	4.42	28.49	2.07
2	{Employment=An employee & Dishwasher=1&Game console=0}	{ECPC2}**	4.24	26.81	1.95
3	{Employment=An employee & freezer=0&cook=yes}	{ECPC2}**	4.48	26.70	1.94
4	{Employment=An employee & Game console=0&cook=yes}	{ECPC2}**	4.39	24.79	1.80
5	{Employment=An employee &Education=third level &cook=yes}	{ECPC2}**	4.27	24.78	1.80
6	{Employment=An employee &Age of respondent=36~56&cook=yes}	{ECPC2}*	5.29	24.44	1.78
7	{Employment=An employee& Bedrooms=4}	{ECPC2}**	4.15	23.23	1.69
8	{Employment=An employee& Occupants=both adults and children}	{ECPC2}**	4.39	23.10	1.68
9	{Occupants=both adults and children &cook=yes}	{ECPC2}**	4.45	22.12	1.61
10	{Social class=C& Dishwasher=1&Electric heater=0}	{ECPC2}**	5.05	22.11	1.61
	Rules related to ECPC3				
1	{Education=secondary level & Dishwasher=1& cook=yes}	{ECPC3}**	7.46	35.53	1.44
2	{Education=secondary level& Dishwasher=1& Electric heater=1}	{ECPC3}**	7.85	33.63	1.36
3	{freezer=1&cook=yes}	{ECPC3}**	10.64	31.49	1.27
4	{TV>21inch=2+&cook=yes}	{ECPC3}**	7.76	31.39	1.27
5	{Bedrooms=3&cook=yes}	{ECPC3}**	8.90	31.09	1.26
6	{Education=secondary level &Occupants=all adults}	{ECPC3}**	7.79	31.02	1.26
7	{Occupants=all adults & cook=yes}	{ECPC3}**	11.15	30.81	1.25
8	{Social class=DE& cook=yes}	{ECPC3}**	7.76	30.79	1.24
9	{Dwelling age=30~75years&cook=yes}	{ECPC3}**	7.75	30.68	1.24
10	{ Dwelling age=30~75years&Occupants=all adults}	{ECPC3}**	7.43	30.31	1.23
	Rules related to ECPC4				
1	{Cook=no}	{ECPC4}**	7.61	24.56	1.38
2	{Electric cooker=0}	{ECPC4}**	5.74	24.55	1.38
3	{Occupants=live alone}	{ECPC4}**	5.63	23.91	1.34
4	{Freezer=0&TV's(21 inch+)=1}	{ECPC4}*	5.65	21.86	1.23
5	{Dwelling type=bungalow house}	{ECPC4}*	5.56	20.58	1.15
	Rules related to ECPC5				
1	{Age of respondent=55+&Laptop=0&cook=yes}	{ECPC5}**	4.24	23.42	1.78
2	{Age of respondent=55+&Electric cooker=1&Laptop=0}	{ECPC5}**	4.48	22.44	1.71
3	{Employment =retired & cook=yes}	{ECPC5}**	4.84	22.39	1.70
4	{Employment =retired& Occupants=all adults}	{ECPC5}**	4.57	22.16	1.69
5	{Age of respondent=55+&Income=<30k}	{ECPC5}**	4.06	22.09	1.68
6	{Income=<30k&Electric cooker=1}	{ECPC5}**	4.18	20.72	1.58
7	{Social class=DE & Occupants=all adults}	{ECPC5}**	4.42	19.78	1.51
8	{Social class=DE & cook=yes}	{ECPC5}**	4.96	19.69	1.50
9	{Occupants=all adults &Laptop=0}	{ECPC5}**	4.75	19.68	1.50
10	{Dishwasher=0&cook=yes}	{ECPC5}**	3.97	19.13	1.46
	Rules related to ECPC6				
1	{Education=third level & TV>21inch=1}	{ECPC6}**	1.38	7.63	1.71
2	{Age of respondent=36~55&Education=third level}	{ECPC6}**	1.41	7.32	1.65
3	{Age of respondent=36~55&TV>21inch=1}	{ECPC6}**	1.44	6.71	1.51

4	{Age of respondent=36~55&freezer=0}	{ECPC6}**	1.38	6.57	1.48
5	{Employment=An employee & freezer=0}	{ECPC6}**	1.50	6.52	1.46
6	{Employment=An employee& Education=third level}	{ECPC6}**	1.47	6.50	1.46
7	{Occupants=both adult and children}	{ECPC6}**	1.71	6.21	1.40
8	{Social class=C&TV>21inch=1}	{ECPC6}**	1.35	6.09	1.37
9	{Electric heater=1}	{ECPC6}*	1.47	6.06	1.36
10	{freezer=0&TV>21inch=1}	{ECPC6}**	1.53	5.93	1.33

* P-value<5% ; ** P-value<1%

4.5. ARM results on sub-item set

It can be seen from Table 4 that some rules of ECPC1 have the same LHS with rules of ECPC4, including '*Occupants=live alone*', '*Cook=no*' and '*Electric cooker=0*'.In order to investigate how electricity consumption behavior changes with different HCs, we try to control the above three HCs and explore the impact of other HCs on ECP in spring. A sub-item set was generated by picking out the customers satisfying the above three constraints from the *item set0*, as such, we can obtain a sub-item set (called *sub-item set1*) containing 221 customers. Then the enhanced Apriori was performed on the *sub-item set1* in spring, the parameters were set as same as *item0*. The rules related to ECPC1 and ECPC4 were sorted by *Lift* values and shown in Table 5.We can find that households living in semi-detached houses with dishwashers but no TV>21 inch are more likely to use electricity like ECPC1. On the contrary, households living in bungalow houses with TVs>21inch but no dishwasher present significant associations with ECPC4. The above examples indicate that the ECP are influenced by multiple HCs. We can just roughly distinguish the ECPC1 and ECPC4 from other clusters depending on the *occupants*, *cook by electricity or not* and *number of electric cooker* three HCs. However, it is difficult to determine whether a household's ECP belongs to ECPC1 or ECPC4 only depending on the above three HCs. More HCs are needed to explain the variances of electricity consumption behaviors.

Table 5 Interesting rules associated with ECPC1 and ECPC4 discovered from sub-item set1

Rules	LHS	RHS	Sup(%)	Conf(%)	Lift
	Rules related to ECPC1				
1	{Double-glazing=all &TV>21inch=0}	{ECPC1}**	12.22	55.10	1.35
2	{Dwelling type=semi-detached }	{ECPC1}**	14.03	50.00	1.23
3	{Dishwasher=1}	{ECPC1}**	16.74	49.33	1.21
	Rules related to ECPC4				
1	{Double-glazing=all &freezer=0&TV>21inch=1}	{ECPC4}**	9.50	46.67	1.61
2	{CFL=None& Dishwasher=0}	{ECPC4}**	9.05	41.67	1.44
3	{Dwelling type= bungalow house}	{ECPC4}**	9.95	40.74	1.40

** P-value<1%

4.6. ARM results for summer,autumn and winter

The ARM results for summer, autumn and winter are shown in Fig. 10 by grouped matrix-based visualization technique [39]. *R*-extension package *arulesViz* [40] is chosen to visualize the grouped matrix by a balloon plot with antecedent groups as columns and consequents as rows. The meaning of this figure can be illustrated as follows. Taking the rules in summer as examples, the group of most interesting rules according to *Lift* measure are shown in the top-left corner of the plot. There are 2 rules which contain 'Employment=retired' and up to 4 other items in the antecedent and the consequent is 'ECPC5'.

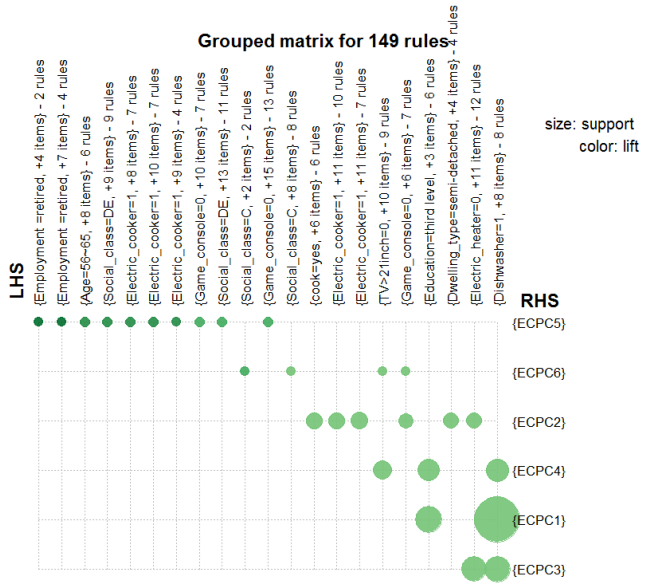
5. Discussions

5.1. Summary

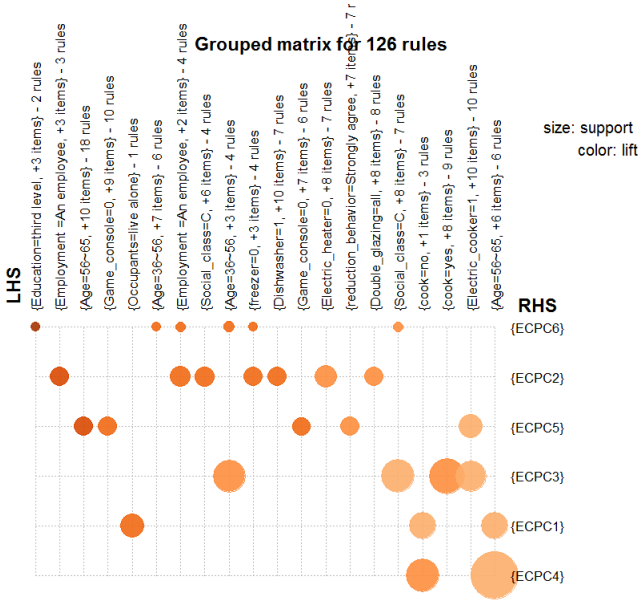
Significant variation on the number of association rules can be found for four categories of HCs. There are more than half of rules associating socio-demographic and cooking related factors with ECPs, which indicates that these two kinds of HCs have major impacts on ECPs. No significant relation is found between attitudes related factors and ECPs for all seasons. The effect of

dwelling factors varies across seasons.

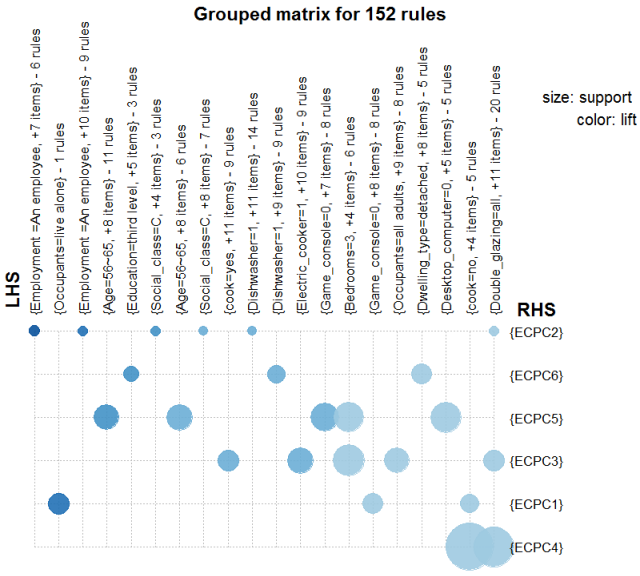
Employment status shows the strongest association with residential ECPs in spring, summer and winter, whereas no significant relation is found between *sex* and ECPs. Other socio-demographic factors such as *occupants, age of respondent, education level* and *social class* present great significant association across all seasons. However, *income* has relatively small effects on ECPs. One reason might be that more than half of customers refuse to provide their income information.



(a) summer



(b) autumn



(c) winter

Fig. 10 Grouped matrix-based visualization for rules in summer, autumn and winter. A balloon plot with antecedent groups as columns and consequents as rows was used. The bigger the size of the balloon is, the larger the aggregated support value will be. The deeper the colors of the balloons are, the larger the values of the aggregated *Lift* in the group with a certain consequent will be. The number of antecedents and the most important (frequent) items in the group are displayed as the labels for the columns.

Dwelling factors can hardly affect ECPs in summer. For other seasons, associations, but not strong, can be found between some dwelling factors such as *dwelling type, bedrooms and dwelling age* and ECPs. In particular, *double glazing* is found to be associated with ECPs in both autumn and winter probably due to its relationship to heat loss in cold weather.

Regarding the appliances and heating factors, cooking related factors such as *cook, electric cooker and dishwasher* present

stronger associations with ECPs than others. Entertainment appliances such as *Game console* and *TV>21 inch* also have important impacts on ECPs. No relationship is found between some wet appliances such as *washing machines*, *tumble dryers*, *water pump and Immersion* and ECPs.

It is a little strange that the attitudes related factors almost have no effect on ECPs. These factors may affect the total value of electricity consumption to some extent. However, the occurring times of electricity usage for cooking usually are fixed during the whole day for most customers because they prepare breakfast, lunch and dinner at almost the same time every day and hardly be influenced by attitudes.

The functioning mechanism of different HCs on ECPs can be discussed from two different time scales: intra-daily and seasonal variation. For intra-daily scale, the rough daily load shape is mainly related to two HCs: *employment status* and *occupants* that basically determine the number of demand peaks and the time of occurrence. For example, householders working as employees usually exhibit two demand peaks in morning and evening like ECPC2 and ECPC6, nevertheless these ECPs are unlikely to appear in retired households. Actually, *employment status* and *occupants* have significant impacts on ECPs by affecting the number of people staying at home and the staying duration. Other socio-demographic factors such as *social class*, *income, education level* and *age of respondent* are highly correlated with *employment status*, thus they may play indirect roles in affecting the electricity consumption behaviors. Similarly, the dwelling factors such as *bedrooms* and *dwelling age* are correlated with *occupants*, thereby these factors also can be considered as indirect impact factors. The magnitude and occurring time of demand peak are mainly determined by cooking related factors (e.g. *cook*, *dishwasher*) and entertainment appliances (e.g. *TV>21inch, game console*). Those households cooking by electricity tend to show larger magnitudes of demand peaks during the time of meal such as ECPC3 and ECPC5.

For seasonal time scale, it can be summarized from the variations of three different aspects: daytime, occupancy and temperature. First, the variation of daytime affects ECPs through the change of work-rest schedule (e.g. commuter time, getting up and sleeping). Regarding the change of *occupants*, although it is unlikely to be changed in a short term (e.g. one month or one year), the number of people staying at home during daytime may vary across seasons due to various holidays (e.g. the summer holiday for children and Christmas holiday). So those households with more than one person (especially with children) are more likely to change ECP across seasons like ECPC2, 3, 5 and 6. Whereas those people living alone usually do not change their ECPs across seasons like ECPC1 and 4. Third, the temperature variation has an impact on ECP through the change of heating appliances usage such as *electric heater*. Dwelling factors such as *double glazing* and *insulated walls* are highly correlated with the insulation ability of house, thus these factors only play indirect roles in affecting ECPs through their impacts on heating appliance usage under low temperature conditions. Finally, it is understandable that the ECP in spring is very similar with the corresponding ECP in autumn, because they exhibit similar characteristics in the above three aspects.

5.2. Applications of the study

The proposed approach in this paper provides a basic framework to identify the key HCs of residential ECPs. Several potential applications based on the proposed approach and findings can be discussed from the following four aspects.

For customers, the knowledge obtained by this study can increase the benefits that they could receive through more targeted, customized and comprehensive energy services. For example, more targeted efficiency campaigns, such as the social interaction based electricity reduction program [41], can be provided to specific customer groups showing similar ECPs, which might help to make efficiency-related topics interesting and therefore increase customer engagement .These interesting campaigns will not only decrease customers' electricity bills to save money but also can reduce the carbon emission at the meantime.

For utilities, the proposed approach can help them to find suitable customers who have the potential of peak shaving to participate in demand peak reduction schemes. The targeted services such as new tailor-made tariff schemes can be provided based on the knowledge of ECPs to match the customer's specific lifestyles. On the basis of the findings, utilities can offer directed electricity savings advice to their customers. It has been widely indicated that customized services and electricity bill savings advice can improve the satisfaction of customer. Furthermore, the findings of this study are helpful for improving the accuracy of load forecasting and baseline load estimation [42, 43]. In recent years, the ECPs of residential customers have changed significantly with the increasing penetration of distributed solar PV systems. Although some PV power forecasting methods [44-46] can be used to predict the output power of solar PV systems, it is still difficult to accurately predict the load of an individual customer because some HCs with significant associations with ECPs such as *employment status*, *occupants* and *cook* are not considered in the traditional load foresting models. Hence, these factors can be used as predictors to further improve the accuracy of the current load forecasting models or estimate the ECP for new customer.

For policy makers, the studies based on the approach proposed in this paper can help them to get under the skin of the energy using habits, so as to better understand where do the influences on electricity consumption come from and then figure out how to form new policies to influence and lead (through legislative prohibition or financial incentives or disincentives) people into desired paths of using electricity more efficiently and friendly.

In terms of the methodological implications, the case study was conducted using the dataset from Ireland, however the approach proposed in this paper can be applied to the datasets from any other region. It is noted that conducting survey to acquire customer information especially detailed information like the survey used in this paper is a high cost and time-consuming work. Hence, based on the findings of this study, more efforts should be made on those important HCs (e.g. socio-demographic factors) instead of the other HCs. In addition, the methodology can also be used to reveal the inverse association relations from ECPs to HCs by setting the rules containing ECP variable as LHS and the survey variables as RHS.

6. Conclusions

This paper proposed a quantitative analysis approach using smart metering data and survey information to identify the most significant HCs affecting residential ECPs. Clustering methods (i.e. adaptive DBSCAN and K-means) were used to extract the TECP of each customer and group customers showing similar TECPs together. Association analysis method (i.e. enhanced Apriori algorithm) was used to reveal the relationship between HCs (i.e. including dwelling characteristics, socio-demographic, appliances and heating and attitudes towards energy) and residential TECPs. The main findings of this paper are listed as follows:

1) Socio-demographic factors except for *sex* show strong significant associations with ECPs for all seasons.

2) The effect of dwelling factors on ECPs varies across seasons. Except for summer, associations, but not strong, can be found between these factors and ECPs in other seasons.

3) Cooking related factors such as *whether cook by electricity* show strong significant associations with ECPs. However, no relationship is found between some wet appliances such as *washing machines* and ECPs.

4) Attitudes related factors almost have no effect on ECPs.

5) Those households with more than one person (especially with children) are more likely to change ECP across seasons.

At last, we argued the values of relevant studies based on the proposed approach and analyzed the applicability of the methodology. However, some important HCs such as income, floor area and air-conditioning were not investigated entirely or included in this study due to the limitation of data source. The ECPs of weekend days were not investigated either due to the space limitation of the paper. The future works of this research are listed as follows:

1) Testing the proposed approach on more datasets and explore the relations between more HCs and ECPs under various scenarios.

2) Air-conditioning system plays an important role in demand side management. In order to understand the ECP of air-conditioning and how it relates to other HCs, we plan to use Non-intrusive load monitoring techniques [47] to separate the electricity consumption of air-conditioning system from the total electricity consumption and then use the proposed approach to analyze the relation between air-conditioning use pattern and HCs.

3) The proposed approach can be combined with some other machine learning methods (e.g. support vector machine and K-Nearest Neighbor [48]) to identify the HCs from smart meter data.

Appendix

Given a set of observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$, where each observation is a *d*-dimensional real vector, K-means clustering aims to partition the *n* observations into *K* ($\leqslant$ n) sets $S=(S_1, S_2, ..., S_K)$ so as to minimize the within-cluster sum of squares. Formally, the objective is to find

$$\arg\min_{S} \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in S_k} \left\| \boldsymbol{x} - \boldsymbol{C}_k \right\|^2$$

Where $\boldsymbol{C}_k$ is the cluster centroid *k*. The pseudo-code of K-means algorithm is shown in Algorithm 1.

Algorithm 1: K-means algorithm

Input: K //the Number of Clusters

Select K data points randomly as initial cluster centroids $\boldsymbol{C}=[\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_K]$ // Initialize cluster centroids
Repeat //iteration $\boldsymbol{i}$
for all points $x_j \in$ N **do** //Calculate the distance between x_j and the cluster centroid $\boldsymbol{C}_w$
for all data points $\boldsymbol{C}_w \in \boldsymbol{C}$ **do**
calculate $||x_j-\mathbf{C}_w||$
end for
select the minimum $||x_j-\mathbf{C}_w||$ and label x_j as $\boldsymbol{C}_w$ //label x_j
end for
$C_w = 1/N_w \sum_{x \in C_w} x$ //Recalculate the centroid for each cluster
Until there is no change for each cluster or meets the maximum time of iteration
Return clustering result

The pseudo-code of adaptive DBSCAN is shown in Algorithm 2.

Algorithm 2: Adaptive DBSCAN

Input: DB, distFunc, eps, MinPts, Δeps, Np
do
C←0 //Cluster counter
for each point P **in** database DB **do**
if label(P)≠undefined **then**
continue //Previously processed in inner loop
end if
Neighbors N=RangeQuery(DB, distFunc, P, eps) // Find neighbors
if |**N**|<MinPts **then**// Density check
label(P)=Noise //Label as Noise
continue
end if
C=C+1 // next cluster label
label(P)=C // Label initial point
Seed set S=N\{P} // Neighbors to expand
for each point Q **in** S **do**
if label(Q)=Noise **then**
label(Q)=C // Change Noise to border point
end if
if label(Q) ≠undefined **then**
continue //previously processed
end if
label(Q)=C //Label neighbor
Neighbors N= RangeQuery(DB, distFunc, P, eps) //Find neighbors
if |**N**|≥minPts **then**
S= S∪N //Add new neighbors to seed set
end if
end for
end for
Count the number of core points Np
if Np>0 **then**
break
else eps=eps+Δeps
end if
While(1)
Store the current value of eps and the indexes of the core points
end

Define function RangeQuery

```
        Input:DB, distFunc, Q, eps
37:     Neighbors=empty list
38:     for each point P in database DB do  //Scan all points in the database
39:      if distFunc(Q,P)≤ eps then  //Compute distance and check epsilon
40:         Neighbors=Neighbors∪{P}   //Add to result
41:      end if
42:     end for
43:     return Neighbors
44:     end
```


Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (grant No. 51577067), the Beijing Natural Science Foundation of China (grant No. 3162033), the Hebei Natural Science Foundation of China (grant No. E2015502060), the State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (grant Nos. LAPS18008), the Open Fund of State Key Laboratory of Operation and Control of Renewable Energy & Storage Systems (China Electric Power Research Institute) (No. 5242001600FB), the Fundamental Research Funds for the Central Universities (No. 2018QN077). This work was also supported by the U.S. Department of Energy under Contract No. DE-AC36-08-GO28308 with the National Renewable Energy Laboratory. M. Shafie-khah and J. P. S. Catalão acknowledge the support by FEDER funds through COMPETE 2020 and by Portuguese funds through FCT, under Projects SAICT- PAC/0004/2015 - POCI-01-0145-FEDER-016434, POCI-01-0145- FEDER-006961, UID/EEA/50014/2013, UID/CEC/50021/2013, and UID/EMS/00151/2013, and also funding from the EU 7th Frame- work Programme FP7/2007–2013 under GA No. 309048.

REFERENCES

[1] Electricity information 2017 Overview, IEA. [Online]. Avaliable: http://www.iea.org/publications/freepublications/publication/ElectricityInformation2017Overview.pdf, accessed Mar.12, 2018.

[2] Department of Energy & Climate Change (DECC). United Kingdom Housing Energy Fact File 2013. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/345141/uk_housing_fact_file_2013.pdf, accessed Mar.26, 2017.

[3] Smart Meters Market by Type (Electric, Water, and Gas), Application (Commercial, Residential, and Industrial), Technology (Automatic Meter Reading and Advanced Metering Infrastructure), and by Region - Global Forecasts to 2022, Market Research. [Online]. Avaliable: http://www.rnrmarketresearch.com/smart-meters-market-by-type-smart-electric-meters-smart-water-meters-smart-gas-meters-by-end-user-industrial-commercial-and-residential-region-americas-asia-pacific-europe-row-tren-market-report.html, accessed Mar.12, 2018.

[4] Ville Uusitalo, Sanni Väisänen, Eero Inkeri, Risto Soukka. Potential for greenhouse gas emission reductions using surplus electricity in hydrogen, methane and methanol production via electrolysis. Energy Convers Manage 2017;134:125-134.

[5] CRU Announces Delivery Plan for Smart Meters in Ireland, Commission for Regulation of Utilities. [Online]. Avaliable: https://www.cru.ie/2017/07/28/post-2/, accessed Mar.12, 2018.

[6] B. Yildiz, J.I. Bilbao, J. Dore, A.B. Sproul. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. Appl Energy 2017;208:402-427.

[7] R. Li, Z. Wang, C. Gu, F. Li, and H. Wu. A novel time-of-use tariff design based on Gaussian Mixture Model. Appl Energy 2016;162:1530-1536.

[8] A. Lahouar, J. Ben Hadj Slama. Day-ahead load forecast using random forest and expert input selection, Energy Convers Manage 2015;103:1040-1051.

[9] Thair S. Mahmoud, Daryoush Habibi, Mohammed Y. Hassan, et al. Modelling self-optimised short term load forecasting for medium voltage loads using tunning fuzzy systems and Artificial Neural Networks. Energy Convers Manage 2015,106:1396-1408.

[10] Saboori H, Hemmati R, Abbasi V. Multistage distribution network expansion planning considering the emerging energy storage systems. Energy Convers Manag 2015;105:938–945.

[11] Wolisz H, Punkenburg C, Streblow R, Müller D. Feasibility and potential of thermal demand side management in residential buildings considering different developments in the German energy market. Energy Convers Manag 2016;107:86–95.

[12] F. Wang, H. Xu, T. Xu et al. The values of market-based demand response on improving power system reliability under extreme circumstances. Appl Energy 2017;193:220-231.

[13] F. Wang, L. Zhou, H. Ren et al. Multi-objective Optimization Model of Source-Load-Storage Synergetic Dispatch for Building Energy System Based on TOU Price Demand Response. IEEE Trans. Ind Appl 2017;54:1017-1028.

[14] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012;42:68-80.

[15] G. Chicco, I.-S. Ilie. Support vector clustering of electrical load pattern data. IEEE Trans Power Syst 2009; 24:1619–1628.

[16] G. Chicco, O. Ionel, R. Porumb. Electrical load pattern grouping based on centroid model with ant colony clustering. IEEE Trans Power Syst 2013; 28:1706-1715.

[17] Q. Chen, F. Wang, B.M. Hodge et al. Dynamic price vector formation model based automatic demand response strategy for PV-assisted EV charging station. IEEE Trans. Smart Grid 2017;8:2903-2915.

[18] J. Kwac, J. Flora, R. Rajagopal. Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid 2014;5:420-430.
[19] Azimi R, Ghayekhloo M, Ghofrani M. A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. Energy Convers Manag 2016;118:331-344.
[20] G. Chicco, R. Napoli, F. Piglione. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21:933-940.
[21] Jin CH, Pok G, Lee Y, Park HW, Kim KD, Yun U, et al. A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting. Energy Convers Manag 2015;90:84–92.
[22] Granell R, Axon CJ, Wallom DCH. Clustering disaggregated load profiles using a Dirichlet process mixture model. Energy Convers Manag 2015;92:507-516.
[23] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014;135:461-471.
[24] Fintan McLoughlin, Aidan Duffy, Michael Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190-199.
[25] Beckel C, Sadamori L, Staake T, Santini S. Revealing household characteristics from smart meter data. Energy 2014;78:397-410.
[26] Joaquim L. Viegas, Susana M. Vieira, R. Melício et al. Classification of new electricity customers based on surveys and smart metering data. Energy 2016;107:804-817.
[27] Demirhan H. The problem of multicollinearity in horizontal solar radiation estimation models and a new model for Turkey. Energy Convers Manag 2014;84:334-345.
[28] Irish Social Science Data Archive. Data from the Commission for Energy Regulation (CER)-smart metering project. [Online]. Avaliable: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/, accessed Dec.10, 2017.
[29] Gesche Huebner, David Shipworth, Ian Hamilton et al. Understanding electricity consumption: A comparative contributionof building factors, socio, demographics, appliances, behaviours and attitudes. Appl Energy 2016;177:692-702.
[30] The Commission for Energy Regulation et al. CER National Smart Metering Programme Status Update [Online]. Avaliable: https://www.cer.ie/docs/001021/CER16126%20NSMP%20Information%20Paper%20(1).pdf, accessed Dec.10, 2017.
[31] Seasons in Ireland, Seasons of the Year. [Online]. Avaliable: https://seasonsyear.com/Ireland, accessed Jan.12, 2018.
[32] M. Ester, H.P. Kriegel, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 2nd ACM SIGKDD, Portland, Oregon, 1996,pp. 226-231.
[33] DBSCAN, wikipedia. [Online]. Avaliable: https://en.wikipedia.org/wiki/DBSCAN, accessed Jan.12, 2018.
[34] Agrawal, R., Imielinski, T., Swami, A. Mining associations between sets of items in massive databases. In Proc.of the 1993 ACM-SIGMOD Int'l Conf. on Management of Data, 207-216.
[35] Izwan Nizal Mohd. Shaharanee, Fedja Hadzic, Tharam S. Dillon. Interestingness measures for association rules based on statistical validity. Knowledge Based Syst.2011;24:386-392.
[36] Rand Wilcox, Chapter 9 - Correlation and Tests of Independence, In Statistical Modeling and Decision Science, Academic Press, 2017, Pages 485-516, Introduction to Robust Estimation and Hypothesis Testing (Fourth Edition).
[37] Liu Bing, Wynne Hsu, Yiming Ma. Pruning and summarizing the discovered associations. In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press 1999;125-134.
[38] Bayardo, R. , R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. Data Mining and Knowledge Discovery 2000;4(2/3):217–240.
[39] Hahsler M, Chelluboina S.Visualizing association rules in hierarchical groups. In 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011). The Interface Foundation of North America.
[40] Michael Hahsler, Sudheer Chelluboina.Visualizing association rules: introduction to the R-extension package arulesViz. [Online]. Avaliable: https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf, accessed Jan.12, 2018.
[41] Fei Wang, Liming Liu, Yili Yu, et al. Impact Analysis of Customized Feedback Interventions on Residential Electricity Load Consumption Behavior. Energies 2018;11:1-22.
[42] Liang X, Hong T, Shen GQ. Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Appl Energy 2016;179:247–260.
[43] F. Wang, K. Li, C. Liu et al. Synchronous Pattern Matching Principle Based Residential Demand Response Baseline Estimation: Mechanism Analysis and Approach Description. IEEE Trans. Smart Grid 2018;pp:1-1.
[44] F. Wang, Z. Zhen, C. Liu et al. Image phase shift invariance based cloud motion displacement vector calculation method for ultra-short-term solar PV power forecasting. Energy Convers Manag 2018;157:123-135.
[45] F. Wang, Z. Mi, S. Su et al. Short-Term Solar Irradiance Forecasting Model Based on Artificial Neural Network Using Statistical Feature Parameters. Energies 2012;5:1355-1370.
[46] Y. Sun, F. Wang, B. Wang et al. Correlation Feature Selection and Mutual Information Theory Based Quantitative Research on Meteorological Impact Factors of Module Temperature for Solar Photovoltaic Systems. Energies 2017;10:1-20.
[47] Su S, Yan Y, Lu H, et al. Non-intrusive load monitoring of air conditioning using low-resolution smart meter data. 2016 IEEE Int Conf Power Syst Technol 2016:1–5.
[48] Wang F, Zhen Z, Wang B et al. Comparative Study on KNN and SVM Based Weather Classification Models for Day Ahead Short Term Solar PV Power Forecasting. Appl Sci 2017;8:1-23.