Communications for Statistical Applications and Methods
2022, Vol. 29, No. 2, 225–238

https://doi.org/10.29220/CSAM.2022.29.2.225
Print ISSN 2287-7843 / Online ISSN 2383-4757

Clustering non-stationary advanced metering infrastructure data

Donghyun Kang[a], Yaeji Lim[1,a]

[a]Department of Applied Statistics, Chung-Ang University, Korea

Abstract

In this paper, we propose a clustering method for advanced metering infrastructure (AMI) data in Korea. As AMI data presents non-stationarity, we consider time-dependent frequency domain principal components analysis, which is a proper method for locally stationary time series data. We develop a new clustering method based on time-varying eigenvectors, and our method provides a meaningful result that is different from the clustering results obtained by employing conventional methods, such as K-means and K-centres functional clustering. Simulation study demonstrates the superiority of the proposed approach. We further apply the clustering results to the evaluation of the electricity price system in South Korea, and validate the reform of the progressive electricity tariff system.

Keywords: advanced metering infrastructure (AMI), clustering, non-stationary data, progressive electricity tariff system, time-dependent frequency domain principal components analysis

1. Introduction

Smart grid appears as a key term in the context of optimizing energy efficiency. One of the integral facets of the smart grid is advanced metering infrastructure (AMI), which enables two-way communication meters and the availability of electricity consumption data at a higher frequency. By 2021, UK energy suppliers will have installed 50 million smart meters, and 22.5 million smart meter installations are planned in South Korea.

Therefore, various studies have analyzed AMI data in relation to energy consumption in households. Krishna *et al.* (2015) proposed an anomaly detection method based on principal component analysis (PCA) and density-based spatial clustering of applications with noise to verify the integrity of the smart meter measurements. Glasgo *et al.* (2017) used direct metering methods, non-intrusive load monitoring methods, and statistical methods to inform estimates of device-level energy consumption in the United States, and Chalmers *et al.* (2019) examined electricity usage, behaviors and changes in habits from smart meter records.

Various clustering methods have been applied to characterize electricity consumption data and interpret the load pattern. Kwac *et al.* (2014) proposed an electricity customer segmentation methodology that classified households according to extracted features, and Xu *et al.* (2015) applied the hierarchical K-means method to a large-scale AMI dataset. Blakely *et al.* (2019) proposed a spectral clustering approach using a voltage time series produced from AMI equipment to validate and correct customer electrical phase labels. Shin *et al.* (2011) forecasted load pattern based on the clustering

[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail:yaeji.lim@gmail.com

Published 31 March 2022 / journal homepage: http://csam.or.kr

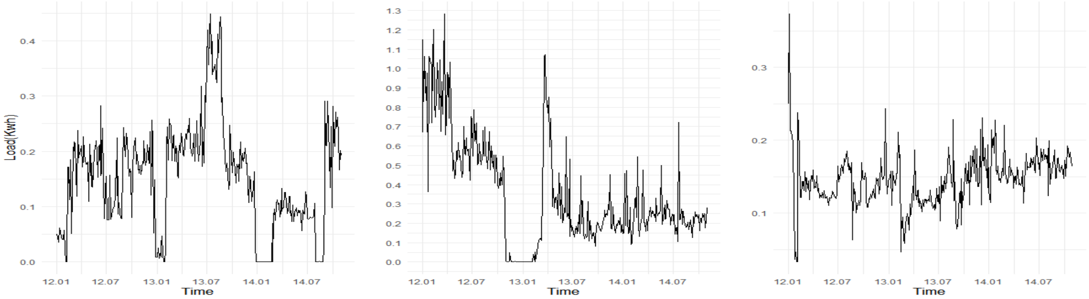

Figure 1: *Three day averaged load pattern from randomly selected households in South Korea from January, 2012 to October, 2014.*

and classification algorithms of temporal data-mining techniques. Further analyses in this field can be found in Chicco *et al.* (2004) and Romero *et al.* (2011). However, most existing studies simply apply conventional clustering methods to the electricity consumption data. Therefore, the non-stationary characteristic of the AMI data has not been considered.

In this paper, we develop a data-adaptive clustering method that may reflect this characteristic of the AMI data and also provide a new perspective on electricity consumption compared to the existing clustering methods. The proposed method is based on the functional PCA (FPCA), which is commonly used for data reduction in functional data (Van Der Linde, 2008; Silverman and Ramsay, 1997). Although concepts and methods in functional data analysis may be robust for serial dependence, they have been developed for independent observations. However, time series data such as economy and energy consumptions, does not support this assumption. To solve this problem, Brillinger (2001) first suggested frequency domain PCA (FDPCA), which considers the correlation in time. FDPCA is also called a dynamic PCA (DPCA), and has been studied in various fields (Salvador *et al.*, 2003; Mina and Verde, 2007).

However, FDPCA is based on the spectral representation of a stationary signal. Stationarity, which means statistical characteristics such as mean and variance do not change over time, is a very important assumption in time series analysis. However, in real data analysis, the stationarity assumption is rarely satisfied (Hamilton, 1989; Azadeh *et al.*, 2010).

Figure 1, which depicts the load pattern from randomly selected households in South Korea from January, 2012 to October, 2014, illustrates the non-stationary characteristic of the data.

To alleviate the stationarity assumption, Ombao and Ho (2006) proposed a time-dependent FDPCA for the locally stationary data. They formed a small block around each time point, and therefore, the multi-channel signal was almost stationary in that time block, which defined the time-varying spectral density matrix. Their method extracted time-varying spectral features of a multi-channel signal. Thus, we expect that the time-dependent FDPCA is a proper starting point to cluster the non-stationary time series data.

In this paper, we propose a time-dependent FDPCA based clustering method for AMI data that is non-stationary time series data. The clustering results are further applied to validate the reform of the progressive electricity tariff system in South Korea. The South Korean government has planned to apply differentiated tariff rates by season and hour, which involves charging higher rates during peak seasons/hours and lower rates during non-peak seasons/hours, and was scheduled to start in July, 2021. We apply the reformed system to each cluster group, and analyze the effect of the reform on household energy usage.

The remainder of this paper is organized as follows: In Section 2, we describe the proposed time-dependent frequency domain principal component clustering method and also provide a practical algorithm. Simulation results are presented in Section 3, and real AMI data analysis are illustrated in Section 4. Finally, the concluding remarks are presented in Section 5.

2. Time-dependent frequency domain principal component clustering

Here, we first briefly review the FDPCA and time-dependent FDPCA, and then explain the proposed clustering method based on the time-dependent FDPCA.

2.1. Time-dependent frequency domain principal component analysis (FDPCA)

Suppose that we have p-dimensional stationary time series $\mathbf{X}(t) = [X_1(t), \ldots, X_p(t)]'$ with zero mean. The corresponding spectral representation is

$$\mathbf{X}(t) = \int_{(-\pi,\pi)} \mathbf{A}(\omega) \exp(i\omega t) d\mathbf{Z}(\omega), \ t \in \mathbb{Z}, \tag{2.1}$$

where $\mathbf{A}(\omega)$ is a time-invariant transfer function matrix and $\mathbf{Z}(\omega)$ is a stochastic process with orthonormal increments. The corresponding spectral density matrix of $\mathbf{X}$(t) is defined to be

$$\mathbf{f}(\omega) = \mathbf{A}(\omega)\overline{\mathbf{A}(\omega)}',$$

where $\overline{\mathbf{A}(\omega)}'$ denotes the transpose of the complex conjugate of $\mathbf{A}(\omega)$, with $-\infty < \omega < \infty$.

We now intend to approximate $\mathbf{X}(t)$ using a q-variate process $\mathbf{J}(t) = [J_1(t), \ldots, J_q(t)]'$ $(q \leqslant p)$, defined by

$$\mathbf{J}(t) = \Sigma_{l=-\infty}^{\infty} \mathbf{c}'(t-l)\mathbf{X}(\mathbf{l}),$$

where $\mathbf{c}(u)$ is the $p \times q$ valued filter and absolutely summable.
Then, we can reconstruct $\mathbf{X}(t)$ from $\mathbf{J}(t)$ by defining

$$\hat{\mathbf{X}}(t) = \Sigma_{l=-\infty}^{\infty} \mathbf{b}(t-l)\mathbf{J}(\mathbf{l}),$$

where $\mathbf{b}(u)$ is the $p \times q$ filter and absolutely summable.
To obtain $\hat{\mathbf{X}}(t)$, minimize the mean square approximation error

$$E[\overline{\{\mathbf{X}(t) - \Sigma_{l=-\infty}^{\infty} \mathbf{b}(t-l)\mathbf{J}(\mathbf{l})\}}' \{\mathbf{X}(t) - \Sigma_{l=-\infty}^{\infty} \mathbf{b}(t-l)\mathbf{J}(\mathbf{l})\}]. \tag{2.2}$$

Then, the minimum solution is

$$\mathbf{c}(\omega) = [\mathbf{V}_1(\omega), \cdots, \mathbf{V}_q(\omega)] = \overline{\mathbf{b}(\omega)}',$$

where $\mathbf{V}_1(\omega), \ldots, \mathbf{V}_q(\omega)$ are eigenfunctions of the spectral density matrix, $\mathbf{f}(\omega)$, and correspond to the eigenvalues, $v_1(\omega) > \cdots > v_q(\omega)$, which are arranged in decreasing magnitude (for a detailed description of the FDPCA, see Theorem 9.3.1 in Brillinger (2001) and Chapter 7 in Shumway *et al.* (2000)).

As mentioned above, the stationarity assumption is crucial in the FDPCA. Ombao and Ho (2006) divided the time series into blocks to apply FDPCA to the locally stationary process. Therefore, the time series within the block appears stationary. Then, define $\mathbf{X}(t)$ as

$$\mathbf{X}(t) = \int_{(-\pi,\pi)} \mathbf{A}\left(\frac{t}{T}, \omega\right) \exp(i\omega t) d\mathbf{Z}(\omega), \ t \in \mathbb{Z}, \tag{2.3}$$

where T is the length of $\mathbf{X}(t)$ and its transfer function matrix, $\mathbf{A}$, now depends on time. The time-varying spectral density matrix is defined as

$$\mathbf{f}(u,\omega) = \mathbf{A}(u,\omega)\overline{\mathbf{A}(u,\omega)}',$$

where $u \in [0,1]$ denotes the rescaled time point and $\omega \in (-\pi,\pi)$.

Ombao and Ho (2006) conducted the FDPCA in the neighborhood of a particular rescaled time $u^* \in [0,1]$. Under the locally stationary process, the spectrum $\mathbf{f}(u^*,\omega)$ is stable and can be estimated by using the time series around $t^* = [u^*T]$, where $[\cdot]$ denotes the greatest integer function. Let $M = 2^J$ be the number of observations in a neighborhood for some positive integer J. Then, $\{\mathbf{X}(t^* - (M/2) + 1,)\ldots,\mathbf{X}(t^* + M/2)\}$ become time blocks and Fourier coefficients are obtained as

$$\mathbf{d}\left(u^*,\omega_k\right) = \frac{1}{\sqrt{M}}\sum_{t=t^*-\frac{M}{2}+1}^{t^*+\frac{M}{2}}\mathbf{X}(t)\exp\left(-i\omega_k t'\right), \tag{2.4}$$

where $t' = t^* - M/2$ and $\omega_k = 2\pi k/M,\ k = -M/2+1,\ldots,M/2$. Then, the periodogram matrix and its smoothed version by the rectangular kernel are computed as

$$\begin{aligned}\mathbf{I}(u^*,\omega_k) &= \frac{1}{2\pi}\mathbf{d}(u^*,\omega_k)\overline{\mathbf{d}(u^*,\omega_k)}',\\ \hat{\mathbf{f}}(u^*,\omega_k) &= \frac{1}{H(u^*)}\sum_{j=-h(u^*)}^{h(u^*)}\mathbf{I}(u^*,\omega_{k+j}),\end{aligned} \tag{2.5}$$

where $H(u^*) = 2h(u^*) + 1$ and $h(u^*)$ is a positive integer.

Finally, the time-varying eigenvalues and eigenvectors are obtained by applying eigenanalysis to the matrix $\hat{\mathbf{f}}(u^*,\omega_k)$. The first q eigenvalues, $\hat{\xi}^1(u^*,\omega_k),\ldots,\hat{\xi}^q(u^*,\omega_k)$, are interpreted as the *time-varying spectrum*. The corresponding eigenvectors are denoted as $\hat{\mathbf{\Xi}}^{\mathbf{1}}(u^*,\omega_k),\ldots,\hat{\mathbf{\Xi}}^{\mathbf{q}}(u^*,\omega_k)$.

Compared to the eigenvalues, $v_1(\omega) > \cdots > v_q(\omega)$, and the corresponding eigenfunctions,
$\mathbf{V}_1(\omega),\ldots,\mathbf{V}_q(\omega)$ obtained from the FDPCA, the time-dependent FDPCA provides time-varying eigenvalues and eigenfunctions.

Therefore, the dimensions of the eigenvalues and eigenvectors in the time-dependent FDPCA are higher than those of the FDPCA. The eigenvalue, $\hat{\xi}^j(u^*,\omega_k)$, in the time-dependent FDPCA can be represented by a time-frequency map, and the $p \times q$ dimensional matrices are obtained for each u^* and ω_k in the eigenfunction, $\hat{\mathbf{\Xi}}^{\mathbf{j}}(u^*,\omega_k)$ for $j = 1,\ldots,q$.

2.2. Clustering

To cluster the data based on the time-varying FDPCA, we define weighted time-varying eigenvectors as follows. For each time point, u^*, and frequency, ω_k, the jth eigenfunction, $\hat{\mathbf{\Xi}}^{\mathbf{j}}(u^*,\omega_k)$, is multiplied by the corresponding eigenvalue at the same time point and frequency, $\hat{\xi}^j(u^*,\omega_k)$. Then, we compute the sum of all frequencies. That is, the p-dimensional weighted time-varying eigenvector, $\mathbf{W}^j(u^*)$, is defined as,

$$\mathbf{W}^j(u^*) := \sum_{\omega_k}\left[\hat{\xi}^j(u^*,\omega_k)\times\hat{\mathbf{\Xi}}^j(u^*,\omega_k)\right], \quad \text{for } j = 1,\ldots,q. \tag{2.6}$$

Now, we perform the conventional K-means clustering (Likas *et al.*, 2003) to the first weighted time-varying eigenvector, $\mathbf{W}^1(u^*)$, for $u^* \in [0,1]$.

2.3. Algorithm

Here, we present a practical algorithm of the proposed method.

Algorithm 1 Time-dependent frequency domain principal component clustering (TFDPC clustering)

```
Inputs:
    $\mathbf{X}(t) = [X_1(t), \ldots, X_p(t)]'$.
    Number of shift $\epsilon$.
    Number of blocks $N$.
    Number of observations in neighborhood $M$.
Initialize:
    $s \leftarrow 1$ be the starting index.
    $e \leftarrow s + M - 1$ be the end index.
for $b = 1$ to $N$ do
    Let $\mathbf{X_t} = [X_1(t), \ldots, X_p(t)]', t = s, \ldots, e$, be the $p$-dimensional time series.
    Calculate Fourier coefficients $\mathbf{d}(u^*, \omega_k)$ as (2.4).
    Construct the smoothed periodogram matrix as follows.
        ○ Calculate $\mathbf{I}(u^*, \omega_k) = \frac{1}{2\pi}\mathbf{d}(u^*, \omega_k)\overline{\mathbf{d}(u^*, \omega_k)}'$.
        ○ $\hat{\mathbf{f}}(u^*, \omega_k) = \frac{1}{H(u^*)} \sum_{j=-h(u^*)}^{h(u^*)} \mathbf{I}(u^*, \omega_{k+j})$.
    Applying eigenanalysis to the matrix $\hat{\mathbf{f}}(u^*, \omega_k)$.
    Obtain the first $q$ eigenvalues, $\hat{\xi}^1(u^*, \omega_k), \ldots, \hat{\xi}^q(u^*, \omega_k)$, and the corresponding eigenvectors, $\hat{\mathbf{\Xi}}^{\mathbf{1}}(u^*, \omega_k), \ldots, \hat{\mathbf{\Xi}}^{\mathbf{q}}(u^*, \omega_k)$.
    $s \leftarrow s + \epsilon$
    $e \leftarrow e + \epsilon$
end for
Computing the first $p$-dimensional weighted time-varying eigenvector, $\mathbf{W}^1(u^*)$, defined as (2.6).
Performing $K$-means clustering to $\mathbf{W}^1(u^*)$.
```

3. Simulation

Here, we conduct simulation studies to confirm the performance of the proposed method.

3.1. Simulation setting

We generate 50 curves from K different processes, $\{X_j^{(k)}(t)\}$, for $k = 1, \ldots, K$. Each curve length is 100 non-stationary time series, and $[X_{(1)}^1(t), \ldots, X_{50}^{(1)}(t), \ldots, X_1^{(K)}(t), \ldots, X_{50}^{(K)}(t)]$ are input curves for each clustering method.

The simulation settings are similar to that of Fryzlewicz and Ombao (2009) with small modifications.

1. Case 1 – Nonstationary autoregressive processes with abruptly changing parameters

$$X_j^{(k)}(t) = \phi_1^{(k)}(t)X_j^{(k)}(t-1) + \phi_2^{(k)}(t)X_j^{(k)}(t-2) + \epsilon_j(t), \qquad \text{for } j = 1, \ldots, 50,$$

where $\epsilon_j(t) \sim N(0, 1)$, and the parameters are defined as follows:

(a) Two groups ($K = 2$) with

$$\phi_1^{(1)}(t) = \begin{cases} 0.6, & t = 1, \ldots, 50, \\ -0.6, & t = 51, \ldots, 100, \end{cases} \qquad \phi_1^{(2)}(t) = \begin{cases} -0.9, & t = 1, \ldots, 50, \\ 0.9, & t = 51, \ldots, 100, \end{cases}$$
$$\phi_2^{(1)}(t) = -0.6, \qquad \phi_2^{(2)}(t) = -0.81.$$

(b) Two groups ($K = 2$) with

$$\phi_1^{(1)}(t) = \begin{cases} 0.6, & t = 1, \ldots, 33, \\ -0.6, & t = 34, \ldots, 66, \\ 0.6, & t = 67, \ldots, 100, \end{cases} \qquad \phi_1^{(2)}(t) = \begin{cases} -0.9, & t = 1, \ldots, 33, \\ 0.9, & t = 34, \ldots, 66, \\ -0.9, & t = 67, \ldots, 100, \end{cases}$$
$$\phi_2^{(1)}(t) = -0.6, \qquad \phi_2^{(2)}(t) = -0.81.$$

(c) Three groups ($K = 3$) with

$$\phi_1^{(1)}(t) = \begin{cases} 0.9, & t = 1, \ldots, 50, \\ -0.9, & t = 51, \ldots, 100, \end{cases}$$
$$\phi_1^{(2)}(t) = \begin{cases} -0.9, & t = 1, \ldots, 50, \\ 0.9, & t = 51, \ldots, 100, \end{cases}$$
$$\phi_1^{(3)}(t) = \begin{cases} 0.3, & t = 1, \ldots, 50, \\ -0.3, & t = 51, \ldots, 100, \end{cases}$$
$$\phi_2^{(1)}(t) = -0.9, \quad \phi_2^{(2)}(t) = -0.9, \quad \text{and} \quad \phi_2^{(3)}(t) = -0.3.$$

2. Case 2 – Nonstationary autoregressive processes with sinusoidal waves

For $j = 1, \ldots, 50$, curves are generated from the following two processes,

$$\begin{cases} X_j^{(1)}(t) = \phi_1(t)X_j^{(1)}(t-1) + \phi_2(t)X_j^{(1)}(t-2) + \sin t + \epsilon_j(t) \\ X_j^{(2)}(t) = \phi_1(t)X_j^{(2)}(t-1) + \phi_2(t)X_j^{(2)}(t-2) + \cos t + \epsilon_j(t), \end{cases}$$

where $\epsilon_j(t) \sim N(0, \sigma^2)$, and the parameters are

$$\text{(a)} \quad \phi_1(t) = \begin{cases} -0.9, & t = 1, \ldots, 33, \\ 0.9, & t = 34, \ldots, 66, \\ -0.9 & t = 67, \ldots, 100, \end{cases} \quad \text{and} \quad \phi_2(t) = -0.9.$$

$$\text{(b)} \quad \phi_1(t) = \begin{cases} -0.9, & t = 1, \ldots, 25, \\ 0.9, & t = 26, \ldots, 50, \\ -0.9, & t = 51, \ldots, 75, \\ 0.9, & t = 76, \ldots, 100, \end{cases} \quad \text{and} \quad \phi_2(t) = -0.9.$$

The sample curves generated from Case 1-(a) and Case 2-(a) are shown in Figure 2.

3.2. Results

In each case, 100 Monte Carlo simulations are conducted, and the average performances are presented in the following tables. As an evaluation measure, the correct classification rate (CCR) and the adjusted Rand index (aRand) of Hubert and Arabie (1985) are considered. aRand measures the

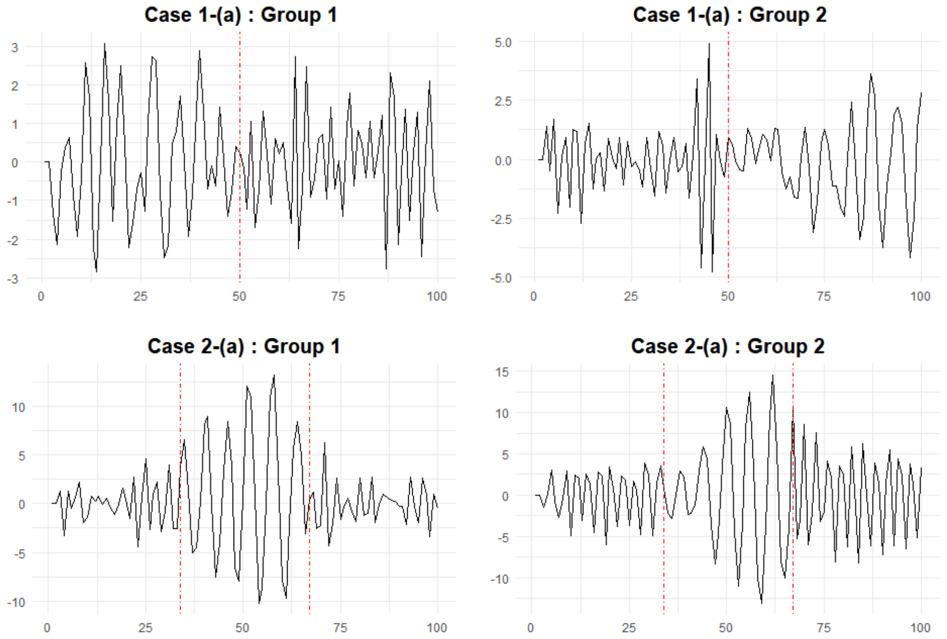

Figure 2: *The sample curves generated from Case 1-(a) (Top) and Case 2-(a) (Bottom) .*

Table 1: Clustering results of Case 1. Bold face indicates the best performance

	Method	CCR	aRand
(a)	TFDPC clustering	**0.715(0.024)**	**0.181(0.041)**
	K-means	0.661(0.046)	0.107(0.051)
	kCFC	0.651(0.077)	0.107(0.092)
(b)	TFDPC clustering	**0.711(0.024)**	**0.175(0.041)**
	K-means	0.661(0.039)	0.104(0.048)
	kCFC	0.604(0.085)	0.063(0.087)
(c)	TFDPC clustering	**0.549(0.027)**	**0.147(0.036)**
	K-means	0.523(0.054)	0.103(0.038)
	kCFC	0.469(0.061)	0.072(0.051)

correspondence between two partitions on how object pairs are classified in the contingency table. A larger aRand value indicates a higher similarity between two partitions.

For comparison, we consider two conventional clustering methods along with the proposed time-dependent FDPC (TFDPC) clustering method. We apply the K-means clustering and K-centres functional clustering method (kCFC) of Chiou and Li (2007) to the simulation datasets. kCFC is also based on the FPCA and can be applied to longitudinal data such as AMI data. The main difference between their method and ours is that we consider the FDPCA, which is more suitable for the time series data that is correlated in time, while kCFC is based on the time domain PCA

From the results, we observe that the proposed TFDPC clustering method works best in Case 1. However, in Case 1-(c), all three methods works poorly and there is no significant difference between the methods.

In Case 2 with small noise variance, $\sigma^2 = 1.5^2$, the CCR of all three methods are close to 0.9, and

Table 2: Clustering results of Case 2. Bold face indicates the best performance

	σ^2	Method	CCR	aRand
(a)	1.5^2	TFDPC clustering	0.901(0.06)	0.653(0.18)
		K-means	**0.921(0.032)**	**0.709(0.107)**
		kCFC	0.871(0.04)	0.551(0.118)
	2.5^2	TFDPC clustering	**0.761(0.072)**	**0.285(0.144)**
		K-means	0.728(0.086)	0.23(0.144)
		kCFC	0.736(0.057)	0.227(0.107)
(b)	1.5^2	TFDPC clustering	0.888(0.036)	0.606(0.106)
		K-means	0.902(0.042)	0.606(0.131)
		kCFC	**0.902(0.041)**	**0.650(0.129)**
	2.5^2	TFDPC clustering	**0.757(0.044)**	**0.265(0.093)**
		K-means	0.691(0.058)	0.15(0.084)
		kCFC	0.598(0.076)	0.052(0.083)

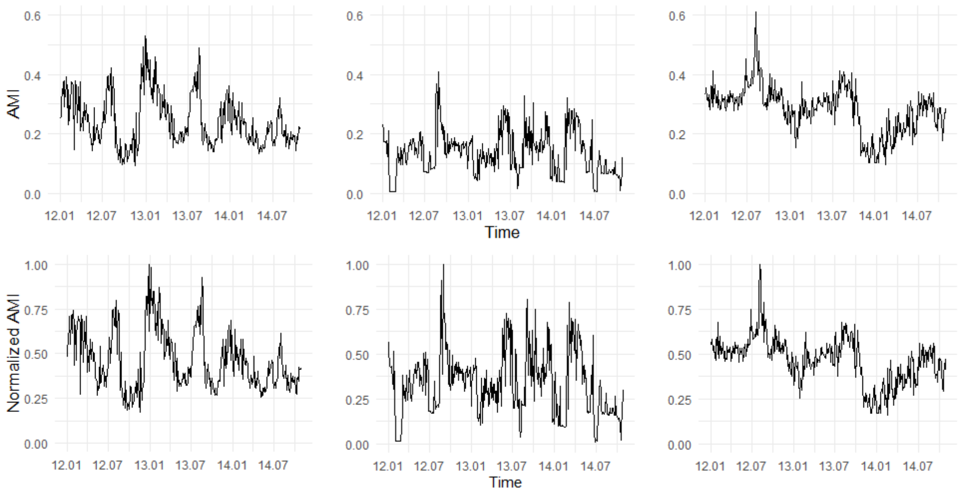

Figure 3: *Randomly selected three AMI data. The top row represents raw AMI datasets, and the bottom row shows the corresponding normalized data from January 2012 to October 2014.*

K-means performs well in both (a) and (b). As the noise variance increases, the performance of all methods deteriorate, but the proposed method does not deteriorate much and works best.

From the simulation results, we conclude that the conventional methods also works well in well-separated cases, but the proposed one works best in some non-stationary autoregressive processes.

4. Advanced metering infrastructure (AMI) data analysis

4.1. Clustering result

The proposed time-varying clustering method is applied in order to assign the load patterns to different clusters. Hourly AMI data were measured for 668 residential customers in Seoul, South Korea from January 1, 2012 to October 31, 2014. The data was sourced from the Korea Electric Power Corporation. For fast computation, we converted hourly data into a three-day averaged time series. Therefore, the number of time points in each AMI dataset is $T = 345$.

We first performed the augmented Dickey–Fuller test to confirm the non-stationarity of the data

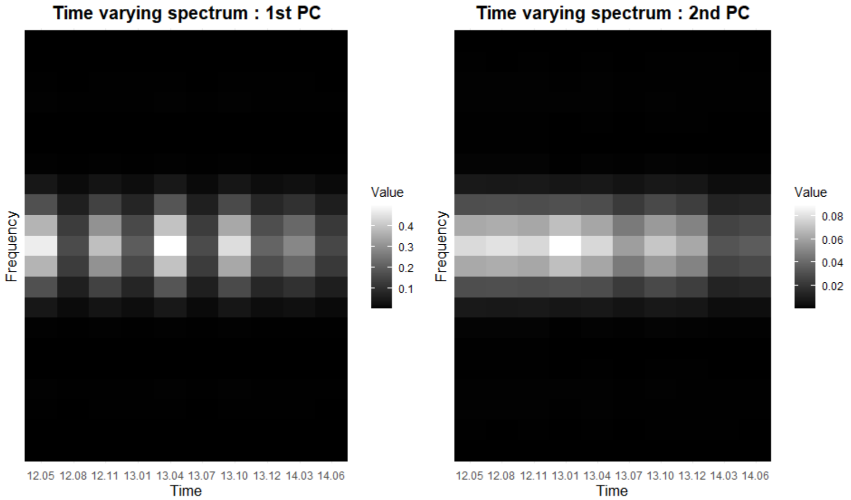

Figure 4: *Time-varying spectrum of the first and second principal components. The darker shade represents low spectral power, whereas the lighter shade represents high spectral power.*

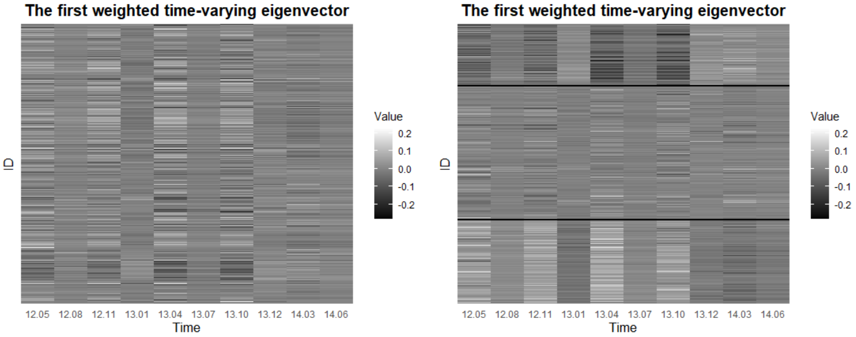

Figure 5: *The first weighted time-varying eigenvector,* $\mathbf{W}^1(u)$*, and ordered weighted time-varying eigenvector according to the* K*-means result. The black solid horizontal lines divide the clustering groups.*

(Said and Dickey, 1984). We find that the null hypothesis cannot be rejected for 189 out of the total 668 customers, implying that some datasets in the AMI time series are non-stationary

Before applying the method, we first normalized each time series. For each AMI dataset, all its measurements are normalized in the range of $[0,1]$ by using its maximum value as the reference power (Xu *et al.*, 2015). The randomly selected three AMI datasets are presented in Figure 3. The top row represents raw AMI data and the bottom row shows the corresponding normalized data. After normalization, the values are distributed from 0 to 1, although the patterns are retained.

We now perform the TFDPC clustering method as follows: We first segment the data into 10 blocks, each with 93 observations and the number of shift is set to 28. Figure 4 plots the first two eigenvalues, $\hat{\xi}^1(u,\omega)$ and $\hat{\xi}^2(u,\omega)$. The center frequencies are brighter than other frequencies in all time points. Further, in the first time-varying spectrum, a relatively higher spectral power is observed periodically.

Now, we compute the first weighted time-varying eigenvector, $\mathbf{W}^1(u)$, as defined in (2.6) and per-

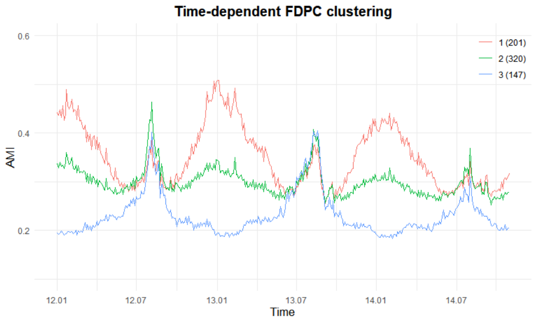

Figure 6: *Time-dependent FDPC clustering result: The averaged AMI time series in each clustering group are plotted. The numbers in the parentheses represent the number of households in each group.*

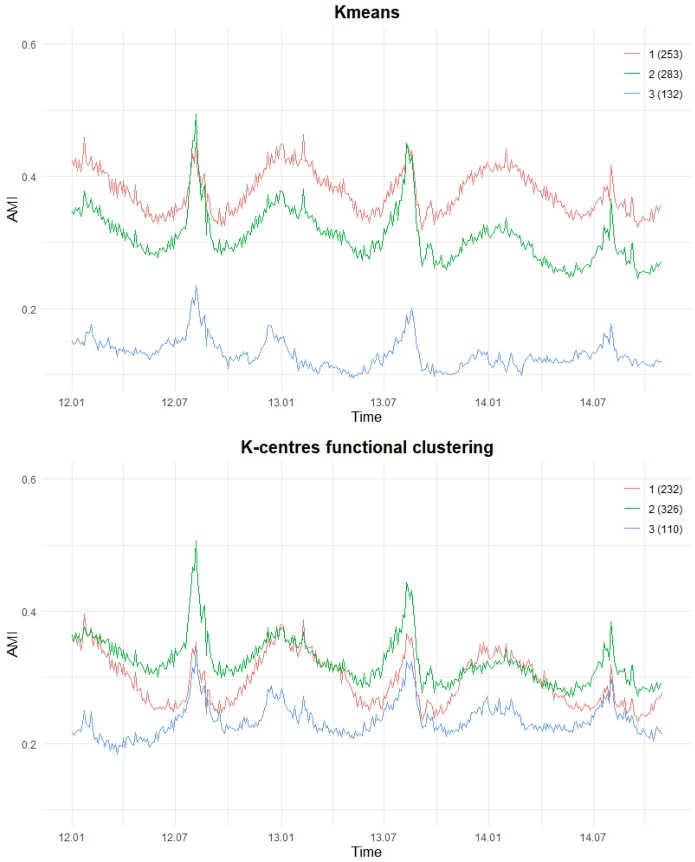

Figure 7: *K-means clustering result (Top) and K-centres functional clustering result (Bottom) : The averaged AMI time series in each clustering group are plotted. The numbers in the parentheses represent the number of households in each group.*

form K-means clustering with $K = 3$. Figure 5 shows the first weighted time-varying eigenvector and re-ordered eigenvector according to the K-means clustering result. All 668 households are clustered

Table 3: Electricity tariff rates for households by bracket before the tariff reform (High-voltage)

	Summer (July-Aug)	
Monthly consumption	Demand charge (won/household)	Energy charge (won/kWh)
1 - 300kWh	730	78.3
301 - 450kWh	1,260	147.3
> 450kWh	6,060	215.6
	Other Seasons	
Monthly consumption	Demand charge (won/household)	Energy charge (won/kWh)
1 - 200kWh	730	78.3
201 - 400kWh	1,260	147.3
> 400kWh	6,060	215.6

Table 4: Electricity tariff rates for households by bracket after tariff reform (Normal Type)

		Summer	Spring/Fall	Winter
		(June - Aug)	(Mar - May) (Sep - Oct)	(Nov - Feb)
Peak load	Time Zone	13 - 17	-	9 - 12
	Charge (won/kWh)	188	-	159
Mid load	Time	9 - 13 & 17 - 23	9 - 23	12 - 23
	Charge (won/kWh)	155	109	138
Off-peak load	Time	23 - 9	23 - 9	23 - 9
	Charge (won/kWh)	82	82	95

in three groups: 201, 320, and 147 households. We observe that the first and the third group display different bright patterns, while the second group contains households that have relatively constant spectral power over time.

Based on the clustering results in Figure 5, we plot the average AMI data for each group in Figure 6. The proposed method clearly classifies households according to the pattern of energy usage. Cluster 1 contains 201 households, and the average AMI time series in this group has relatively high usage during the winter season. By contrast, cluster 3, which contains 147 households, shows a high usage pattern during the summer season and relatively low energy usage in the winter season. Cluster 2 contains 320 households, of which energy usage is high during both the summer and winter, although relatively higher during the summer season.

For comparison, we also apply conventional clustering methods to the normalized AMI data. We apply K-means to the normalized AMI data, and the clustering results are presented in the left plot in Figure 7. Even when we normalize the AMI time series, the K-means method simply classifies the data according to the amount of usage. Therefore, all cluster groups show similar patterns, but with different magnitudes. We also consider kCFC of Chiou and Li (2007). The second plot in Figure 7 shows the kCFC results, and all three groups have relatively higher AMI values during the summer season.

4.2. Application of the electricity price system

The policy of differentiating tariff rates by season and hour involves charging higher rates during peak seasons/hours and lower rates during non-peak seasons/hours. Higher rates are applied in the summer and winter season, and during peak hours. Spring and autumn, as well as sub-peak and off-peak hours are subject to lower tariff rates. Table 3 presents the electricity price system before tariff reform, and Table 4 lists the electricity tariff rates after tariff reform.

Figure 8 presents the box-plot of the electricity prices for each cluster group before tariff reform,

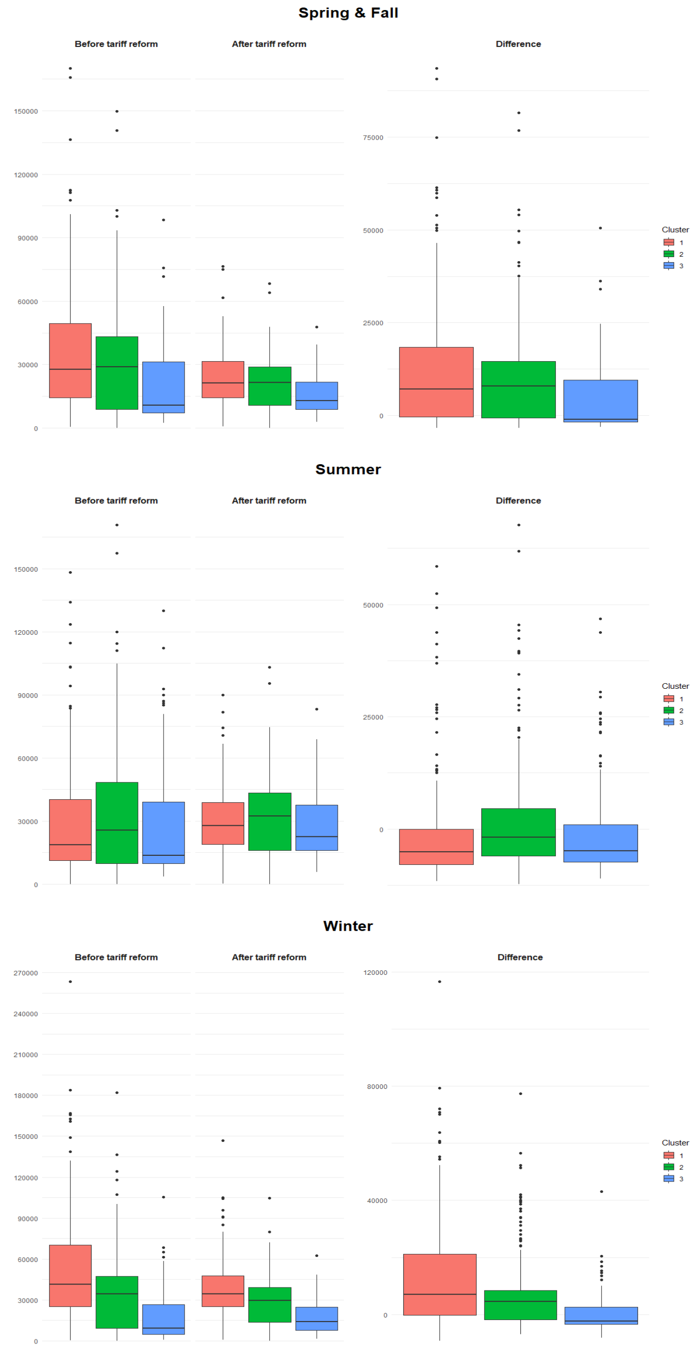

Figure 8: *Box-plot of the electricity prices for each cluster group before the tariff reform, after tariff reform, and the differences.*

after tariff reform, and the differences. As the charges depend on the season, we plot the results according to the season. We observe three outcomes:

1. in summer, the charge on cluster 1, which is the high usage group during the winter season, decreases after tariff reform,
2. in winter, the mean charge on cluster 3, the lowest usage group during the winter, decreases after tariff reform, while it increases for the other cluster groups,
3. in spring and fall, most households' electricity bills increase.

Therefore, we expect that the proposed clustering results can be used to validate the reform of the progressive electricity tariff system in South Korea.

5. Conclusion

In this paper, we propose a new clustering method, the TFDPC clustering method. The proposed method is based on the time-varying FDPCA, and we apply the K-means clustering method to the time-varying eigenanalysis results. We applied the method to residential energy usage and obtained meaningful clustering groups compared to the existing methods. The proposed method was further applied to the electricity price system in South Korea to validate the reform of the progressive electricity tariff system.

From a practical viewpoint, we believe that the proposed method can be extended as follows,

- The clustering results can be applied to characterize residential energy use if more detailed information on residential energy usage is provided. Then, the electricity tariff system can be more precisely customized.
- Other clustering methods can also be applied to the time-varying eigenvector. Comparing these results can offer deeper insights.

Acknowledgment

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2020, National Research Foundation of Korea (NRF) funded by the Korea government (2021R1A2B5B01001790) and Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea (No. 20199710100060).

References

Azadeh A, Saberi M, and Seraj O (2010). An integrated fuzzy regression algorithm for energy consumption estimation with non-stationary data: a case study of Iran, *Energy*, **35**, 2351–2366

Blakely L, Reno MJ, and Feng Wu-chi (2019). Spectral clustering for customer phase identification using AMI voltage timeseries, *2019 IEEE Power and Energy Conference at Illinois (PECI)*, 1–7

Brillinger DR (2001). *Time Series: Data Analysis and Theory*, SIAM

Chalmers C, Hurst W, Mackay M, and Fergus P (2019). Identifying behavioural changes for health monitoring applications using the advanced metering infrastructure, *Behaviour & Information Technology*, **38**, 1154–1166.

Chicco G, Napoli R, Piglione F, Postolache P, Scutariu M, and Toader C (2004). Load pattern-based classification of electricity customers, *IEEE Transactions on Power Systems*,**19**, 1232–1239.

Chiou J-M and Li P-L (2007). Functional clustering and identifying substructures of longitudinal data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 679–699.

Fryzlewicz P and Ombao H (2009). Consistent classification of nonstationary time series using stochastic wavelet representations, *Journal of the American Statistical Association*, **104**, 299–312.

Glasgo B, Hendrickson C, and Azevedo Inês ML (2017). Using advanced metering infrastructure to characterize residential energy use, *The Electricity Journal*, **30**, 64–70.

Hamilton JD (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica: Journal of the Econometric Society*, 357–384.

Hubert L and Arabie P (1985). Comparing partitions, *Journal of Classification*, **2**, 193–218.

Krishna VB, Weaver GA, and Sanders WH (2015). PCA-based method for detecting integrity attacks on advanced metering infrastructure, In *Proceedings of the International Conference on Quantitative Evaluation of Systems*, 70–85.

Kwac JS, Flora J, and Rajagopal R (2014). Household energy consumption segmentation using hourly data, *IEEE Transactions on Smart Grid*, **5**, 420–430.

Likas A, Vlassis N, and Verbeek JJ (2003). The global k-means clustering algorithm, *Pattern Recognition*, **36**, 451–461.

Mina J and Verde C (2007). Fault detection for large scale systems using dynamic principal components analysis with adaptation, *International Journal of Computers Communiations* & *Control*, **2**, 185–194.

Ombao H and Ho MHR (2006). Time-dependent frequency domain principal components analysis of multichannel non-stationary signals, *Computational Statistics* & *Data Analysis*, **50**, 2339–2360.

Silverman BW and Ramsay JO (1997). *Functional Data Analysis*, Amsterdam, Elsevier.

Romero M, Gallego L, and Pavas, Andrés (2011). Estimation of voltage sags patterns with k-means algorithm and clustering of fault zones in high and medium voltage grids, *Ingeniería e Investigación*, **31**, 131–138.

Said SE and Dickey DA (1984). Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika*, **71**, 599–607.

Salvador M, Gallizo JL, and Gargallo P (2003). A dynamic principal components analysis based on multivariate matrix normal dynamic linear models, *Journal of Forecasting*, **22**, 457–478.

Shin JH, Yi BJ, Kim YI, Lee HG, and Ryu KH (2011). Spatiotemporal load-analysis model for electric power distribution facilities using consumer meter-reading data, *IEEE Transactions on Power Delivery*, **26**, 736–743.

Shumway RH, Stoffer DS, and Stoffer DS (2000). *Time Series Analysis and its Applications*, **3**, Springer.

Van Der Linde A (2008). Variational bayesian functional PCA, *Computational Statistics* & *Data Analysis*, **53**, 517–533.

Xu T-S, Chiang H-D, Liu G-Y and Tan C-W (2015). Hierarchical K-means method for clustering large-scale advanced metering infrastructure data, *IEEE Transactions on Power Delivery*, **32**, 609–616.

Received September 15, 2021; Revised October 28, 2021; Accepted October 29, 2021