ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Extracting Urban Water Usage Habits from Smart Meter Data: a Functional Clustering Approach

N. Cheifetz[1], A. Samé[2], Z. Noumir[2], A.-C. Sandraz[1], C. Féliers[1] and V. Heim[3] *

1- Veolia Eau d'Ile de France, 28 Boulevard de Pesaro, F-92751 Nanterre
2- Université Paris-Est, IFSTTAR, COSYS, GRETTIA, F-77447 Marne-la-Vallée
3- Syndicat des Eaux d'Ile de France, 120 Boulevard Saint-Germain, F-75006 Paris

Abstract. Through automated meter reading systems, recent development of smart grids offers the opportunity for an efficient and responsible management of water resources. The present paper describes a novel methodology for identifying relevant usage profiles from hourly water consumption series collected by smart meters located on a water distribution network. The proposed approach operates in two stages. First, an additive time series decomposition model is used in order to extract seasonal patterns from the time series. Then, two functional clustering approaches are used to group the extracted seasonal patterns into homogeneous clusters: a functional version of the well-known K-means algorithm, and a Fourier regression mixture-model-based algorithm. The two clustering strategies are applied to real world data from a smart grid deployed on a large water distribution network in France and a realistic interpretation of the consumption habits is given to each cluster.

1 Introduction

Modern cities deal with increasing populations and climate change, while maintaining adequate water services for consumers. The management of smart cities [1] is now based on automated electronic meters (with fine granularities) that are deployed on a distribution network. First researches in demand patterns classification were made in the electricity field [2]. Most of researches in the water field is focused on demand forecasting [3]. More recently the classification of water demand is addressed by [4] or [5] with a classical version of K-means.

In this paper, the consumption time series collected by smart meters are seen as functions or curves, that is to say functional data[6]. Analyzing smart meter consumption is useful for water utilities in order to develop innovative capabilities in terms of grid management, planning and customer services. The proposed methodology aims to identify automatically major water usage patterns. This is formulated in two consecutive steps: the extraction of seasonal patterns in section 2, and their segmentation in section 3. Two clustering strategies are compared: a functional version of K-means and a dedicated Expectation Maximization (EM) algorithm. The section 4 presents an experimental study and an analysis of the clustering results is given.

[*] This study was part of a previous work between IFSTTAR and Veolia Eau d'Ile de France. This work is now part of the French-German collaborative research project named ResiWater.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

2 Extracting seasonal patterns from time series

Let $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote n univariate time series, where each one of them $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ corresponds to hourly log-consumptions[1] recorded by a single meter. All the time series are assumed to be recorded over the same time grid.

2.1 Fourier-based time series decomposition

The methodology developed in this paper is based on the following classical additive decomposition:

$$y_{it} = f_{it} + x_{it} + d_{it} + \varepsilon_{it}, \tag{1}$$

where

- f_{it} is the global trend of the time series using moving averages.
- x_{it} is the seasonal component, modeling daily and weekly seasonalities. And x_{it} is estimated using a Fourier basis decomposition [3]:

$$\sum_{j=1}^{q_1} \left[\alpha_j^{(1)} \cos\left(\frac{2\pi jt}{24}\right) + \alpha_j^{(2)} \sin\left(\frac{2\pi jt}{24}\right) \right] + \sum_{j=1}^{q_2} \left[\alpha_j^{(3)} \cos\left(\frac{2\pi jt}{168}\right) + \alpha_j^{(4)} \sin\left(\frac{2\pi jt}{168}\right) \right], \tag{2}$$

  where q_1 and q_2 are the numbers of trigonometric terms used to handle the daily and weekly seasonality, and the $\alpha_j^{(*)}$ are to be estimated.
- d_{it} is a component devoted to capture the effect of exceptional public non-working days in France, such as $d_{it} = \sum_{j=1}^{24} \gamma_j \delta_{tj}$ where $\delta_{tj} = 1$ if t corresponds to the hour j of a non-working day and $\delta_{tj} = 0$ otherwise.
- ε_{it} is a centered Gaussian noise.

2.2 Parameters estimation and practical use of the model

First, the trend f_i is estimated using a simple moving average of order 168 (weekly periodicity). Given a couple (q_1, q_2), the coefficients $\alpha_j^{(*)}$ and γ_j are simultaneously identified by performing a multiple linear regression of $(y_{it} - f_{it})$ over the variables $\cos\left(\frac{2\pi jt}{24}\right)$, $\sin\left(\frac{2\pi jt}{24}\right)$, $\cos\left(\frac{2\pi jt}{168}\right)$, $\sin\left(\frac{2\pi jt}{168}\right)$ and δ_{tj}. Selecting the couple (q_1, q_2) is a remaining issue which can be ideally addressed by optimizing a model selection criterion (such as the AIC or BIC). In this paper, several combinations of (q_1, q_2) were tested and the couple $(4, 24)$ has been selected leading to a good compromise between visual representation of seasonal patterns and modeling accuracy. An example of decomposition of a time series is shown in Figure 1. The trend is displayed together with the complete time series while the seasonal component is displayed with the weekly sub-series.

[1] For compliance with the additivity and gaussianity assumptions of this decomposition model, each time series $(y_{it})_{t=1,\ldots,T}$ was replaced by $(\log(y_{it} + \lambda))_{t=1,\ldots,T}$, where λ is a small positive number preventing degeneracy caused by null consumptions.

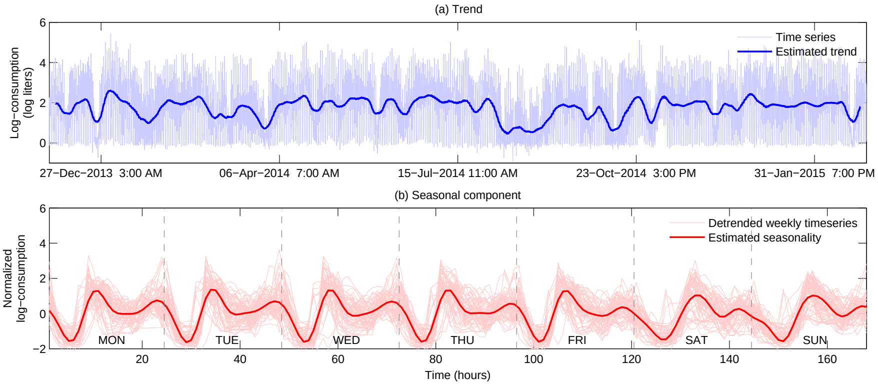


Fig. 1: Extraction of periodic seasonal patterns using a Fourier-based decomposition. The trend is displayed with the complete time series (a) and the seasonal component is displayed with the weekly time series (b).

Due to the periodicity of the series $(x_{it}, \ldots, x_{iT})$ defined by Equation (2), the first terms $m = 168$ are sufficient to characterize the time series. Then, each seasonal pattern is defined by $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, with $m = 168$. It is worth noting that the seasonal patterns $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are standardized [7] such as $\forall i, t, \quad x_{it} \leftarrow \frac{x_{it} - (1/m)\sum_{j=1}^{m} x_{ij}}{\sigma(\mathbf{x}_i)}$, where $\sigma(\mathbf{x}_i)$ is the standard deviation of $\mathbf{x}_i$. The set of normalized seasonal patterns is used as input data for the clustering step that is described in the following section.

3 Clustering seasonal profiles

3.1 Functional clustering based on FPCA

In this subsection, the clustering method is inspired by functional data clustering [6] assuming that data are functions or curves. This approach is is based on the Functional Principal Component Analysis (FPCA) in two consecutive steps:

1. *Smoothing and dimension reduction*: this step consists in applying the classical PCA on the multivariate data obtained by discretizing the functions $(x_1(t), \ldots, x_n(t))$ over the temporal grid $\{1, \ldots, m\}$. In this paper, principal components are selected such as 95% of the data variance is explained.

2. *Clustering*: in this step consists, a classical clustering method is performed on the principal component scores estimated previously. The well-known K-means algorithm is applied using several random initializations and the partition with the lowest intra-cluster inertia is selected.

The resulting functional clustering strategy is called FPCA-KM. The number of cluster K has been selected by minimizing the BIC-like penalized criterion

$BIC(K) = C + \nu_K \log(n)$, where C is the intra-cluster inertia optimized by the K-means algorithm and $\nu_K = Kq$ is the number of parameters to be estimated with q the number of selected principal components.

3.2 Fourier regression mixture model

Inspired by the polynomial regression mixture model [7], this subsection introduces a Fourier regression mixture model, called the FReMix model. Unlike standard vector-based mixture models, the density of each component of the FReMix model is represented by a trigonometric prototype function that is parameterized by regression coefficients and a noise variance. This model therefore assumes that each time series $\mathbf{x}_i$ is distributed according to the following density

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}\right), \tag{3}$$

where $\boldsymbol{\theta} = (\pi_1,\ldots,\pi_K,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K,\sigma_1^2,\ldots,\sigma_K^2)$ is the parameter vector to be estimated. The probabilities π_k are the proportions of the mixture satisfying, $\boldsymbol{\beta}_k \in \mathbb{R}^{2(q_1+q_2)}$ is the coefficient vector of the k-th regression model and $\sigma_k^2 > 0$ is the associated noise variance. The matrix $\mathbf{U} = [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_m]'$ is a $m \times 2(q_1+q_2)$ regression matrix, where the vector $\boldsymbol{u}_t \in \mathbb{R}^{2(q_1+q_2)}$ is defined by $(\forall t = 1,\ldots,m)$

$$\begin{aligned}\boldsymbol{u}_t \quad = \quad & \Big[\cos\left(\frac{2\pi t}{24}\right) \;\; \sin\left(\frac{2\pi t}{24}\right) \;\; \cdots \;\; \cos\left(\frac{2\pi q_1 t}{24}\right) \;\; \sin\left(\frac{2\pi q_1 t}{24}\right) \\ & \cos\left(\frac{2\pi t}{168}\right) \;\; \sin\left(\frac{2\pi t}{168}\right) \;\; \cdots \;\; \cos\left(\frac{2\pi q_2 t}{168}\right) \;\; \sin\left(\frac{2\pi q_2 t}{168}\right)\Big]',\end{aligned}$$

and $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. And the class-specific prototype functions are given by $g_k(t) = \boldsymbol{\beta}_k'\boldsymbol{u}_t$.

Assuming that the n seasonal time series $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ are independent, the parameter vector $\boldsymbol{\theta}$ is estimated by maximizing the following log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}\right), \tag{4}$$

via a dedicated Expectation-Maximization (EM) procedure [8],[7]. The pseudocode can be found in [9]. The number of clusters is selected through the BIC criterion [10] defined by $BIC(K) = -2\mathcal{L}(\widehat{\boldsymbol{\theta}}) + \nu_K \log(n)$, where $\widehat{\boldsymbol{\theta}}$ is the parameter vector estimated by the EM algorithm, and ν_K is the number of free parameters of the model: $\nu_K = 2K(q_1+q_2+1) - 1$.

After estimated the parameter vector $\boldsymbol{\theta}$, a time series partition is obtained by assigning each series $\mathbf{x}_i$ to the cluster having the highest posterior probability

$$\tau_{ik} = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_k, \sigma_k^2\mathbf{I}\right)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_\ell, \sigma_\ell^2\mathbf{I}\right)}. \tag{5}$$

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

4 Experimental study using real data

4.1 Description of the data set

The experimental data set represents the water consumption recorded by a few smart meters deployed on the network of Syndicat des Eaux d'Ile-de-France (SEDIF). The SEDIF is a large association including 150 municipalities which provides drinking water for more than 4 million inhabitants of suburban Paris. The hourly consumption (liter) is measured by $10,233$ meters during 15 months (from Nov-2013 to Mar-2015). Then, the univariate time series are $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, where $n = 10,233$ and the length of each time series $\mathbf{y}_i$ is $T = 11,016$.

4.2 Selecting the number of clusters

The number of clusters for the two methods was selected by running the algorithms with several values of K and then choosing the value which minimizes the BIC criterion. For both methods, the BIC criterion exhibits a decrease continuously while the K value increases. Nevertheless, it can be seen that the variation of BIC is not significant when the number of clusters is above 8. Therefore, the number of clusters is selected such as $K = 8$.

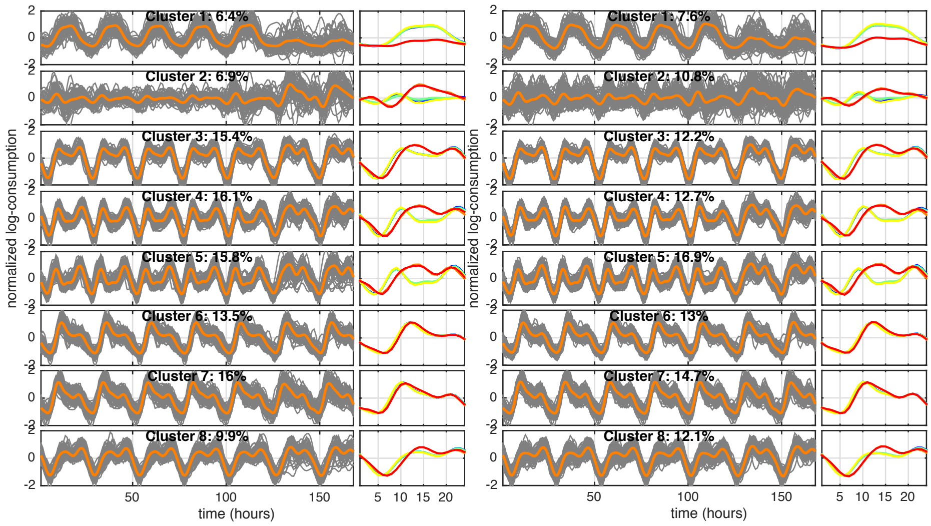
4.3 Results interpretation

Figures 2a and 2b illustrate the results for the two clustering approaches. The consumption profiles are quite similar for the two methods, despite the cluster percentage differences. A qualitative evaluation of the results is performed and the pattern repartition can be explained by the following realistic categories:

- **Office or industrial use**: cluster 1. Active water consumption for the working days and a very low consumption during the weekend.
- **Residential use**: clusters $3, 4, 5$. The temporal dynamic corresponds to customers who take a shower at around 10 AM in the morning. The second peak at around 20 PM corresponds to the home return.
- **Commercial use**: clusters $6, 7, 8$. This category corresponds to a set of customers whose consumption habits are the same during working days and weekends (small businesses, medical centers,...).
- **Noise cluster**: cluster 2. This cluster has the largest variance (groups a set of atypical patterns) and is considered as a noise cluster.

5 Conclusion

A general methodology is introduced for automatically discriminating several water usages and extracting relevant water consumption profiles from time series recorded by smart meters. Eight clusters are identified for two functional clustering methods. The resulting prototypes are quite similar for the two approaches and a realistic category is given for each cluster. More investigations are in progress with the water utility Veolia Eau d'Ile de France in order to refine the clustering results and the methodology will be applied to a larger database.



(a) FPCA-KM

(b) FReMix

Fig. 2: Clustering results obtained with the FPCA-KM strategy (2a) and the FReMix model (2b). The left subfigures represent a weekly view of the clusters with their prototypes displayed in orange. The right subfigures are daily prototypes resulting from the segmentation of the weekly orange curves and colors (from blue to yellow to red) indicate the week day (from Monday to Sunday).

References

[1] T. Nam and T. Pardo. Conceptualizing smart city with dimensions of technology, people, and institutions. In *International Digital Government Research Conference*. ACM, 2011.

[2] Luis Hernández, Carlos Baladrón, Javier M Aguiar, Belén Carro, and Antonio Sánchez-Esguevillas. Classification and clustering of electricity demand patterns in industrial parks. *Energies*, 5(12):5215, 2012.

[3] A. M. De Livera, R. J. Hyndman, and R. D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[4] K Aksela and M Aksela. Demand estimation with automated meter reading in a distribution network. *Journal of Water Resources Planning and Management*, 137(5), 2011.

[5] S. A. McKenna, F. Fusco, and B. J. Eck. Water demand pattern classification from smart meter data. *Procedia Engineering*, 70:1121–1130, 2014.

[6] Julien Jacques and Cristian Preda. Functional data clustering : a survey. *Advances in Data Analysis and Classification*, (3):231–255, 2014.

[7] Scott J Gaffney. *Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models*. PhD thesis, 2004.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1977.

[9] A. Samé, Z. Noumir, N. Cheifetz, A.-C. Sandraz, and C. Féliers. Décomposition et classification de données fonctionnelles pour l'analyse de la consommation d'eau potable (in french). In *EGC conference - Clustering and Co-clustering (CluCo) workshop*, 2016.

[10] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.