Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access
Drinking Water
Engineering and Science
Discussions

Modeling and Clustering Water Demand Patterns from Real-World Smart Meter Data

Nicolas Cheifetz[1], Zineb Noumir[2], Allou Samé[2], Anne-Claire Sandraz[1], Cédric Féliers[1], and Véronique Heim[3]

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

Correspondence to: Nicolas Cheifetz (nicolas.cheifetz@veolia.com)

Abstract.

Nowadays, drinking water utilities need an acute comprehension of the water demand on their distribution network, in order to efficiently operate the optimization of resources, the management of billing and to propose new customer services. With the emergence of smart grids, based on Automated Meter Reading (AMR), a better understanding of the consumption modes is now accessible for smart cities with more granularities. In this context, this paper evaluates a novel methodology for identifying relevant usage profiles from the water consumption data produced by smart meters. The methodology is fully data-driven using the consumption time series which are seen as functions or curves observed with an hourly time step. First, a Fourier-based additive time series decomposition model is introduced to extract seasonal patterns from time series. These patterns are intended to represent the customer habits in terms of water consumption. Two functional clustering approaches are then used to classify the extracted seasonal patterns: the functional version of K-means, and the Fourier REgression Mixture (FReMix) model. The K-means approach produces a hard segmentation and K representative prototypes. On the other hand, the FReMix is a generative model and produces also K profiles as well as a soft segmentation based on the posterior probabilities. The proposed approach is applied to a smart grid deployed on the largest Water Distribution Network (WDN) in France. The two clustering strategies are evaluated and compared. Finally, a realistic interpretation of the consumption habits is given for each cluster. The extensive experiments and the qualitative interpretation of the resulting clusters allow to highlight the effectiveness of the proposed methodology.

1 Introduction

All modern cities need to deal with increasing populations and climate change, while maintaining adequate water services for consumers. Until now, the water or energy consumption reading were traditionally collected once or twice a year on large territories (for example, regions or nations). With the arrival of smart grid meters, this situation has changed and indexes can now be collected automatically with more granularities. The management of smart cities (Giffinger et al., 2007; Nam and Pardo, 2011) is based on automated electronic meters that are deployed on the distribution network and are used to handle billing and customer services. The first researches performed in the area of demand patterns classification belong to the electricity network

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access Drinking Water Engineering and Science Discussions

fields (Irwin et al., 1986; Hernández et al., 2012). Most of the research in water field is focused on demand forecasting (Donkor et al., 2012). Several approaches have been proposed for this purpose, including statistical prevision models (Adamowski, 2008; Blokker et al., 2010). The emergence of smart meters shifts this research to classification of water demand (Aksela and Aksela, 2011). McKenna et al. (2014) proposed a procedure for classification of water demands recorded from smart meters using a Gaussian mixture model not as a clustering strategy but as a feature selection method, and then use the classical K-means algorithm (MacQueen, 1967) for the clustering step.

In various applications, the data to be analyzed are not multivariate observations, but these can be seen as functions or curves either continuous or discrete, namely functional data. Such studies usually refer to Functional Data Analysis (FDA) when data are varying in a continuum and potentially infinite dimensional (Ramsay, 2006; Wang et al., 2015). Example of functional data encompass longitudinal data, responses in medical treatments and objects in video sequences. In the current case of smart meters, each signal is seen as a temporal function and collected intermittently at discrete time points. Analyzing smart meter consumption is useful for water utilities in order to develop innovative capabilities in terms of grid management, planning and customer services. Functional clustering aggregates data mining techniques, which aims to identify homogeneous groups among functional data without using prior knowledge about their group labels (unknown cluster membership). Aiming to analyze household consumption, Cardell-Oliver (2013) introduces a methodology to cluster daily water use signature patterns based on expert rules and a classical K-means. Many functional clustering methods have been developed over the last decade. These methods can usually be separated into two categories: nonparametric methods using specific distances or dissimilarities between curves (Dabo-Niang et al., 2007), and mixture-model-based methods (Samé et al., 2011; Jacques and Preda, 2014). The collected curves can be multivariate leading to a large representation space like in (Cheifetz et al., 2013) for change-point detection based on a specific curve modeling. The approach of the regression mixture model proposed by Gaffney (2003) motivated the focus of this article.

This paper is organized as follows: the overall methodology is described in section 2. This methodology is decomposed in two consecutive steps that is to say the extraction of seasonal patterns from time series in section 3, and the identification of clusters with their profiles in section 4 based on two clustering strategies: a functional version of K-means and a dedicated Expectation Maximization (EM) algorithm. The section 5 introduces the experimental data set, and an analysis of the clustering results is given. Finally, the article ends with a conclusion and some perspectives.

2 General Methodology

The aim of this paper is to identify automatically the major water usage patterns in a set of time series recorded by smart water meters. A multi-step methodology is formulated to address this problematic, as illustrated in Figure 1. The first step consists in extracting the seasonal part of each time series, which represents the habits of water consumption for each meter, using a Fourier-based time series decomposition. Then, these seasonal components are normalized and used as input data by clustering algorithms. Two algorithms are used to classify the functional data into various water usage clusters. The first one consists in using the K-means jointly with the Functional Principal Component Analysis (FPCA) method and the second one is based

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Drinking Water Engineering and Science Discussions
Open Access

on a Fourier regression mixture model recently introduced by Samé et al. (2016). Both the seasonal extraction and clustering approaches will be described in the next sections.

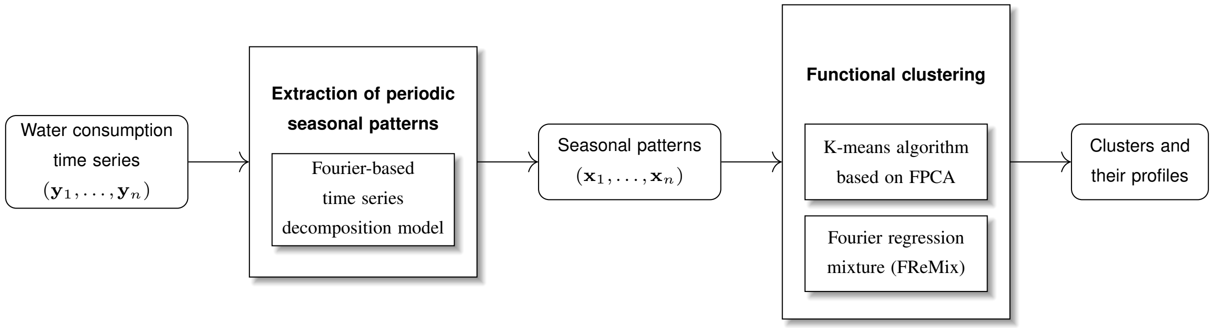


Figure 1. Block diagram describing the global methodology.

3 Extracting seasonal patterns from time series

Let $(\mathbf{y}_1,\ldots,\mathbf{y}_n)$ denote n time series, where each one of them $\mathbf{y}_i=(y_{i1},\ldots,y_{iT})$ corresponds to hourly consumptions recorded by a single water meter, that is to say $\mathbf{y}_i$ is a univariate time series and y_{it} is a real valued scalar. It is implicitly assumed that all the series are recorded over the same time grid indexed by the ordered times $\{1,\ldots,T\}$ for all n curves.

3.1 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 which is modeled in a non parametric way using moving averages (Gourieroux and Monfort, 1997).

- x_{it} is the seasonal component. As the studied water consumption time series are subject to daily and weekly seasonality, a Fourier basis decomposition (De Livera et al., 2011) is formulated:

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

  where q_1 and q_2 are the respective numbers of trigonometric terms used to handle the daily and weekly seasonality, and the $\alpha_j^{(1)},\alpha_j^{(2)},\alpha_j^{(3)},\alpha_j^{(4)}$ are the coefficients to be estimated. This trigonometric modeling has the advantage of requiring considerably less parameters compared to an approach based on dummy variables (De Livera et al., 2011).

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access
Drinking Water
Engineering and Science
Discussions

– d_{it} is a component devoted to capture the effect of exceptional public non-working days in France (eg. January 1^{st}, May 1^{st}, Christmas Day ...). The following decomposition is used: $d_{it} = \sum_{j=1}^{24} \gamma_j \delta_{tj}$, 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.

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. Note that this transformation is used in the same way as the well known Box-Cox transformation (Box and Cox, 1964).

3.2 Parameters estimation and practical use of the model

Given a time series $\mathbf{y}_i$ recorded by a smart meter, the trend f_i is estimated using a simple moving average (Gourieroux and Monfort, 1997; Shumway and Stoffer, 2011). As the daily and weekly periodicities (24 and 168) should be removed from the univariate time series, a centered moving average of order 168 is performed.

After estimating the trend and given a couple (q_1, q_2), the coefficients $\alpha_{1j}, \alpha_{2j}, \alpha_{3j}, \alpha_{4j}$ 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) remains a sensible point which can ideally be addressed by choosing the couple which optimizes a model selection criterion such as the Akaike information criterion (AIC) introduced by Akaike (1974) or the Bayesian information criterion (BIC) introduced by Schwarz (1978). 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 2. The trend is displayed together with the complete time series while the seasonal component is displayed with the weekly sub-series.

From each time series $\mathbf{y}_i$, the model parameters defined by Equation (1) are thus identified, and the periodic seasonal pattern defined by $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, with $m = 168$, is extracted. Due to the periodicity of the series $(x_{it}, \ldots, x_{iT})$ defined by Equation (2), it should be noted that the first terms $m = 168$ are sufficient to characterize the time series. Then, the set of seasonal patterns $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are standardized as suggested by Gaffney (2004)

$$\forall i, t, \quad x_{it} \leftarrow \frac{x_{it} - (1/m)\sum_{j=1}^{m} x_{ij}}{\sigma(\mathbf{x}_i)},$$

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 which will be described in the following section.

It is worth noting that the proposed decomposition can also be used to fill missing values that may occur along the time series. The reconstruction formula is $\widehat{y}_{it} = \widehat{f}_{it} + \widehat{x}_{it} + \widehat{d}_{it}$, where $\widehat{f}_{it}$, $\widehat{x}_{it}$, $\widehat{d}_{it}$ are the estimated components.

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access
Drinking Water
Engineering and Science
Discussions

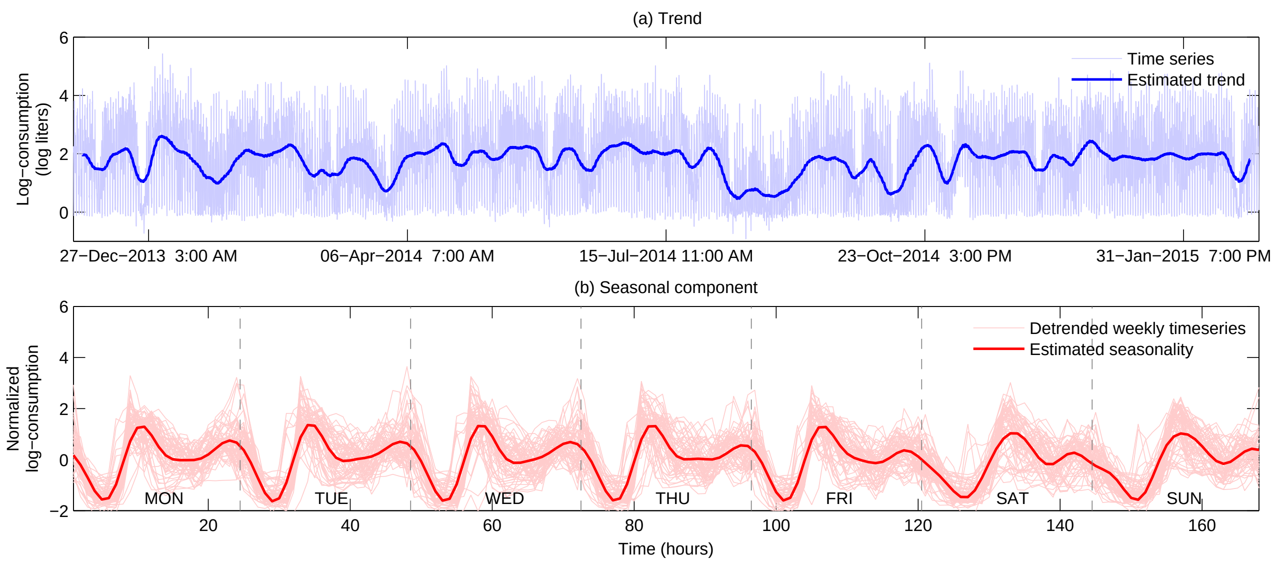


Figure 2. Extraction of periodic seasonal patterns using Fourier-based time series decomposition. The trend is displayed with the complete time series (a) and the seasonal component is displayed with the weekly time series (b).

4 Clustering seasonal profiles

In order to extract relevant usage profiles from water consumption time series, two functional data clustering approaches are considered in this paper: the first one is the functional version of the K-means algorithm and the second one is based on a specific Fourier regression mixture model.

4.1 Functional clustering based on FPCA

In this subsection, the clustering method (Peng and Muller, 2008; Sood et al., 2009) is inspired by functional data analysis (Ramsay and Silverman, 2005; Wang et al., 2015) which assumes that data are functions or curves. This clustering approach is mainly based on Functional Principal Component Analysis (FPCA) and can be summarized into the following two consecutive steps:

1. *Smoothing and dimension reduction*: this step consists in converting the n time series $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ into functional objects $(x_1(t),\ldots,x_n(t))$, and then applying the classical PCA on the multivariate data obtained by discretizing the functions $x_i(t)$ over the temporal grid $\{1,\ldots,m\}$. In this paper, the PCA is directly performed on the seasonal patterns $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ that are based on trigonometric (smooth) functions, and the principal components are selected such as 95% of the data variance is explained.

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

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 (MacQueen, 1967) 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.

The general idea of PCA is to create a small number of uncorrelated variables with maximal variance. The extension of this technique for functional data is proposed in the work of Ramsay and Silverman (2005); Ferraty and Vieu (2006). The FPCA is an efficient tool providing common functional components explaining the structures of individual trajectories.

4.2 Fourier regression mixture model

Inspired by the polynomial regression mixture model formulated by Gaffney and Smyth (1999), this subsection introduces a Fourier regression mixture model, called the FReMix model. The Fourier regression mixture was preferred to polynomial and spline regression mixtures, for its compliance with the modeling adopted in the first step (seasonal pattern extraction). Moreover, Fourier polynomial is a universal approximator of functions and remains a good candidate in modeling clusters whose prototypes are non linear and potentially periodic functions.

4.2.1 Model definition

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. The prototype functions represent the class conditional expectations of $\mathbf{x}_i$. The Fourier regression mixture 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 complete parameter vector. The probabilities π_k are the proportions of the mixture satisfying $\sum_{k=1}^{K}\pi_k = 1$, $\boldsymbol{\beta}_k = (\beta_{k,1},\ldots,\beta_{k,2(q_1+q_2)})' \in \mathbb{R}^{2(q_1+q_2)}$ 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 regression matrix of size $m \times 2(q_1+q_2)$, where the vector $\boldsymbol{u}_t \in \mathbb{R}^{2(q_1+q_2)}$ is defined by ($\forall t = 1,\ldots,m$)

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

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

$$g_k(t) = \sum_{j=1}^{q_1}\left[\beta_{k,2j-1}\cos\left(\frac{2\pi jt}{24}\right) + \beta_{k,2j}\sin\left(\frac{2\pi jt}{24}\right)\right] + \sum_{j=1}^{q_2}\left[\beta_{k,2q_1+2j-1}\cos\left(\frac{2\pi jt}{168}\right) + \beta_{k,2q_1+2j}\sin\left(\frac{2\pi jt}{168}\right)\right]. \tag{4}$$

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access Drinking Water Engineering and Science Discussions

4.2.2 EM algorithm and practical issues

Assuming that the n seasonal time series $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ are independent, the parameter vector $\boldsymbol{\theta}$ is estimated in the same way as for the classical Gaussian mixture model (McLachlan and Krishnan, 2008) and the polynomial regression mixture model (Gaffney and Smyth, 1999), by maximizing the log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}\right), \tag{5}$$

via the Expectation-Maximization (EM) procedure (Dempster et al., 1977; Gaffney and Smyth, 1999; McLachlan and Krishnan, 2008). The pseudo-code can be found in the paper proposed by Samé et al. (2016). As a reminder, the couple $(q_1,q_2) = (4,24)$ is selected in the seasonal pattern extraction step (cf. subsection 3.2). The algorithm is initialized as follows: the initial regression coefficients and variances are obtained by performing a Fourier regression separately on K seasonal series randomly drawn into the data set $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ and the initial proportions of the latent classes are set to $\pi_k = \frac{1}{K}$. This process is repeated 20 times and the parameters with the highest log-likelihood are selected. The number of clusters is selected through the BIC criterion (Schwarz, 1978) 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

$$\tau_{ik} = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}\right)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\mathbf{x}_i; \mathbf{U}\boldsymbol{\beta}_\ell, \sigma_\ell^2 \mathbf{I}\right)}. \tag{6}$$

5 Experimental study using real data

5.1 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. This is the largest drinking WDN in France with about 8,000 km of pipes and more than 750,000 m3 of water produced each day. The consumption is measured hourly (in liter) by $10,233$ meters during 15 months (from Nov-2013 to Mar-2015). The resulting data set is then made of univariate time series $(\mathbf{y}_1,\ldots,\mathbf{y}_n)$, where $n = 10,233$ and the length of each time series $\mathbf{y}_i$ is $T = 11,016$. After the extraction of periodic seasonal patterns (cf. section 3), a new set of time series $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ is built where the length of each seasonal patterns $\mathbf{x}_i$ is $m = 168$. These series are used as input data for the clustering algorithms.

5.2 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. Figure 3 shows the evolution of this criteria for the two clustering algorithms

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access
Drinking Water
Engineering and Science
Discussions

in relation to the number of clusters. 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 $K = 8$.

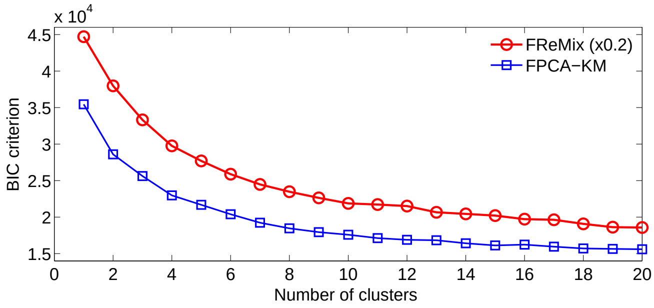


Figure 3. Evolution of the BIC criteria according to the number of clusters.

5.3 Results interpretation

The seasonal time series are classified into $K = 8$ clusters, using functional K-means (FPCA-KM strategy) and Fourier regression mixture (FReMix model) as illustrated respectively by Figures 4a and 4b. For each cluster, the weekly prototype is displayed in orange (sub-figures on the left). Moreover, the right plots of Figures 4a and 4b display the cluster profiles using a daily representation, the colors (from blue to red) indicating the day of the week (from Monday to Sunday). The percentage of input time series belonging to each cluster is also provided.

It can be observed that the consumption profiles are quite similar for the two methods, despite the diffences of the cluster percentages. As no socio-demographic data about customers was available at this stage of the study, a qualitative evaluation of the results is performed and the pattern repartition shown in Figure 4 can be explained by the following realistic categories:

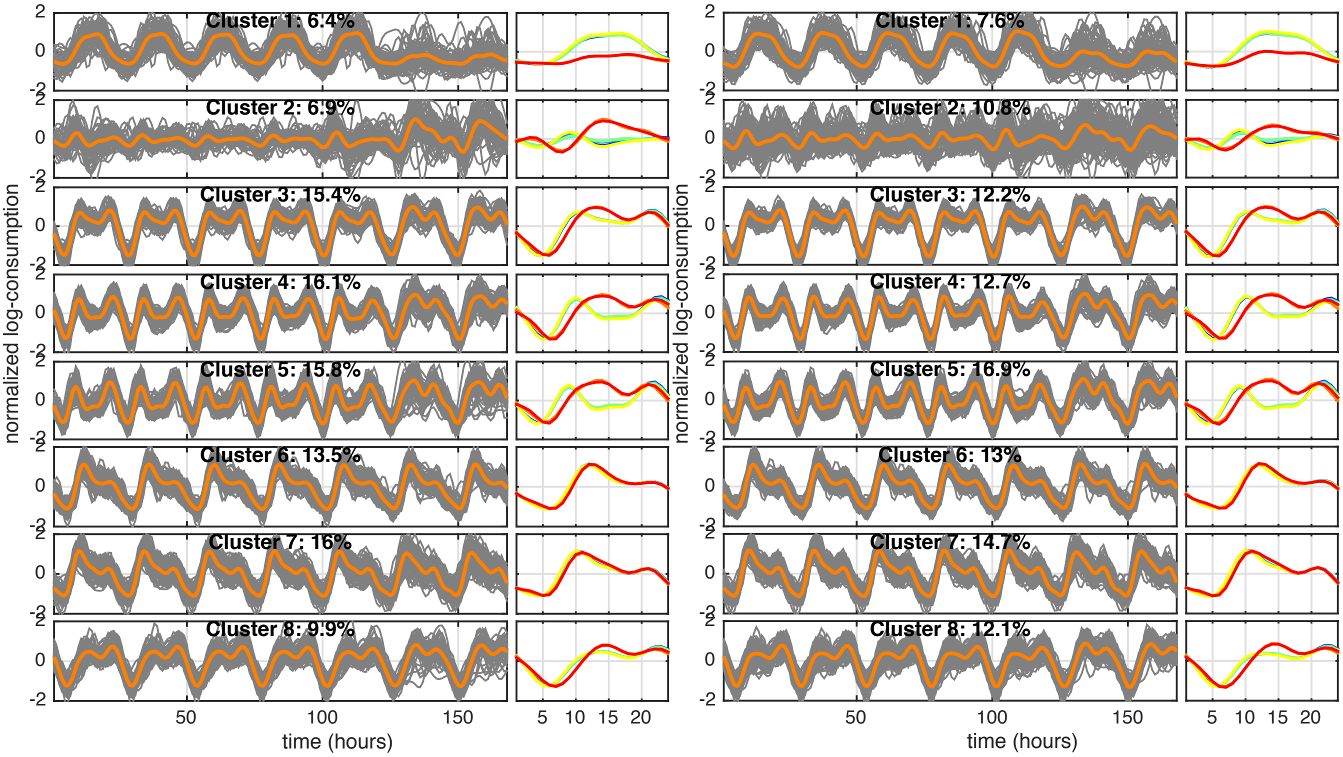
– **Office or industrial use**: cluster 1. One can observe an active water consumption from Monday to Friday (workdays) during the business hours, and a very low consumption during the weekend.

– **Residential use**: clusters $3,4,5$. The temporal dynamic of these clusters corresponds to customers who wake up between 6 AM and 8 AM, take a shower and then go to work. This habit is characterized by a consumption peak around 10 AM in the morning. The other peak, observed in the evening at around 20 PM, corresponds to the return at home. The minimum consumption level between these two peaks can be attributed to persons in households who stay at home during working hours.

– **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. It may correspond, for example, to small businesses or medical centers that stay open every day and have the same daily consumption profile. It should be noticed that clusters $6,7$ differ from the other clusters by their smaller evening peak.

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access
Drinking Water
Engineering and Science
Discussions



(a) FPCA-KM

(b) FReMix

Figure 4. Clustering results obtained with the FPCA-KM strategy (4a) and the FReMix model (4b). For each side, the subfigures on the left represent a weekly view of the clusters with their prototypes displayed in orange. The subfigures on the right are daily prototypes resulting from the segmentation of the weekly orange curves and colors (from blue to yellow to red) indicate the day of the week (from Monday to Sunday).

– **Noise cluster**: cluster 2. This cluster, which has the largest variance, groups a set of atypical patterns which does not match with the other clusters. It can be considered as a noise cluster.

6 Conclusions and perspectives

A general methodology is introduced in this paper for automatically discriminating several water usages and extracting relevant water consumption profiles from time series recorded by smart meters. Considering that the consumption habits are of interest and not the consumption levels, the first step of the method consists in extracting the seasonal part of time series using an additive classical decomposition model. This modeling of the seasonal component is based on a specific Fourier expansion which takes into account daily and weekly periodicities. As this study aims to identify relevant water usage profiles, two functional clustering techniques are used to classify the seasonal patterns extracted from the water consumption time series:

a functional variant of the K-means algorithm and a specific EM algorithm based on a Fourier regression mixture model (FReMix). The FReMix model is richer than the other clustering approach in that the Fourier basis decomposition is fully integrated in the modeling and each cluster is described by its first two moments while the K-means only extracts the mean curves. Furthermore, the K-means produces a hard segmentation while the FReMix creates a soft partition where each cluster membership is weighted by a posterior probability. Eight clusters are then identified for the two clustering methods. The resulting prototypes are quite similar for the two approaches and a realistic category is given 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 proposed methodology is also being applied to a new large scale database.

Acknowledgements. The study presented in the paper 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 ResiWater (http://www.resiwater.eu/) that is funded by the French National Research Agency (ANR; project: ANR-14-PICS-0003) and the German Federal Ministry of Education and Research (BMBF; project: BMBF-13N13690).

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access Drinking Water Engineering and Science Discussions

References

Adamowski, J. F.: Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks, Journal of Water Resources Planning and Management, 134, 119–128, 2008.

Akaike, H.: A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19, 716–723, 1974.

Aksela, K. and Aksela, M.: Demand Estimation with Automated Meter Reading in a Distribution Network, Journal of Water Resources Planning and Management, 137, 456–467, 2011.

Blokker, E. J. M., Vreeburg, J. H. G., and van Dijk J. C.: Simulating Residential Water Demand with a Stochastic End-Use Model, Journal of Water Resources Planning and Management, 136, 19–26, 2010.

Box, G. E. P. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society, Series B, pp. 211–252, 1964.

Cardell-Oliver, R.: Water use signature patterns for analyzing household consumption using medium resolution meter data, Water Resources Research, 49, 8589–8599, 2013.

Cheifetz, N., Samé, A., Aknin, P., De Verdalle, E., and Chenu, D.: A Sequential Testing Procedure for Multiple Change-Point Detection in a Stream of Pneumatic Door Signatures, in: 12th International Conference on Machine Learning and Applications (ICMLA), pp. 117–122, IEEE, 2013.

Dabo-Niang, S., Ferraty, F., and Vieu, P.: On the using of modal curves for radar waveforms classification, Computational Statistics and Data Analysis, 51, 4878–4890, 2007.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D.: Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing, Journal of the American Statistical Association, 106, 1513–1527, 2011.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B, 39, 1–38, 1977.

Donkor, E. A., Mazzuchi, T. A., Soyer, R., and Alan Roberson, J.: Urban water demand forecasting: review of methods and models, Journal of Water Resources Planning and Management, 140, 146–159, 2012.

Ferraty, F. and Vieu, P.: Nonparametric functional data analysis: theory and practice, Springer Science & Business Media, 2006.

Gaffney, S. and Smyth, P.: Trajectory Clustering with Mixtures of Regression Models, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99, pp. 63–72, ACM, New York, NY, USA, 1999.

Gaffney, S. J.: Curve clustering with random effects regression mixtures, in: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003.

Gaffney, S. J.: Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models, Ph.D. thesis, 2004.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., and Meijers, E.: Smart Cities - Ranking of European medium-sized cities, Vienna University of Technology, 2007.

Gourieroux, C. and Monfort, A.: Time series and dynamic models, Cambridge University Press, 1997.

Hernández, L., Baladrón, C., Aguiar, J. M., Carro, B., and Sánchez-Esguevillas, A.: Classification and Clustering of Electricity Demand Patterns in Industrial Parks, Energies, 5, 5215, 2012.

Irwin, G., Monteith, W., and Beattie, W.: Statistical electricity demand modelling from consumer billing data, Generation, Transmission and Distribution, IEE Proceedings C, 133, 328–335, 1986.

Jacques, J. and Preda, C.: Model-based clustering for multivariate functional data, Computational Statistics and Data Analysis, 71, 92–106, 2014.

Drink. Water Eng. Sci. Discuss., doi:10.5194/dwes-2017-19, 2017
Manuscript under review for journal Drink. Water Eng. Sci.
Discussion started: 23 March 2017

Open Access Drinking Water Engineering and Science Discussions

MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281–297, University of California Press, Berkeley, California, 1967.

McKenna, S. A., Fusco, F., and Eck, B. J.: Water Demand Pattern Classification from Smart Meter Data, Procedia Engineering, 70, 1121–1130, 2014.

McLachlan, G. J. and Krishnan, T.: The EM Algorithm and Extensions, Wiley, New York, 2008.

Nam, T. and Pardo, T. A.: Conceptualizing Smart City with Dimensions of Technology, People, and Institutions, in: Proceedings of the 12th Annual International Digital Government Research Conference, ACM, New York, NY, USA, 2011.

Peng, J. and Muller, H.-G.: Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions, Annals of Applied Statistics, 2, 1056–1077, 2008.

Ramsay, J. O.: Functional data analysis, Wiley Online Library, 2006.

Ramsay, J. O. and Silverman, B. W.: Functional Data Analysis, Springer Series in Statistics, Springer, 2nd edn., 2005.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P.: Model-based clustering and segmentation of time series with changes in regime, Advances in Data Analysis and Classification, 5, 301–321, 2011.

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

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

Shumway, R. H. and Stoffer, D. S.: Time series analysis and its applications, Springer, 2011.

Sood, A., James, G. M., and Tellis, G. J.: Functional Regression: A New Model for Predicting Market Penetration of New Products, Marketing Science, 28, 36–51, 2009.

Wang, J.-L., Chiou, J.-M., and Mueller, H.-G.: Review of functional data analysis, arXiv preprint arXiv:1507.05135, 2015.