Mixture of Joint Nonhomogeneous Markov Chains to Cluster and Model Water Consumption Behavior Sequences

Milad Leyli-Abadi, Allou Same, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cedric Féliers, Olivier Chesneau

Mixture of Joint Nonhomogeneous Markov Chains to Cluster and Model Water Consumption Behavior Sequences

MILAD LEYLI-ABADI, ALLOU SAMÉ, and LATIFA OUKHELLOU, Paris-Est Univeristy, IFSTTAR, COSYS, GRETTIA, France
NICOLAS CHEIFETZ, PIERRE MANDEL, and CÉDRIC FÉLIERS, Veolia Eau d'Ile-de-France, France
OLIVIER CHESNEAU, Syndicat des Eaux d'Ile-de-France (SEDIF), France

The emergence of smart meters has fostered the collection of massive data that support a better understanding of consumer behaviors and better management of water resources and networks. The main focus of this paper is to analyze consumption behavior over time; thus, we first identify the main weekly consumption patterns. This approach allows each meter to be represented by a categorical series, where each category corresponds to a weekly consumption behavior. By considering the resulting consumption behavior sequences, we propose a new methodology based on a mixture of nonhomogeneous Markov models to cluster these categorical time series. Using this method, the meters are described by the Markovian dynamics of their cluster. The latent variable that controls cluster membership is estimated alongside the parameters of the Markov model using a novel classification expectation maximization (CEM) algorithm. A specific entropy measure is formulated to evaluate the quality of the estimated partition by considering the joint Markovian dynamics. The proposed clustering model can also be used to predict future consumption behaviors within each cluster. Numerical experiments using real water consumption data provided by a water utility in France and gathered over nineteen months are conducted to evaluate the performance of the proposed approach in terms of both clustering and prediction. The results demonstrate the effectiveness of the proposed method.

CCS Concepts: • **Mathematics of computing** → **Markov networks**; *Time series analysis*; *Cluster analysis*; • **Computing methodologies** → *Unsupervised learning*; *Mixture modeling*.

Additional Key Words and Phrases: non-homogeneous Markov models, categorical time series, clustering, forecasting, water consumption behavior

ACM Reference Format:
Milad Leyli-abadi, Allou Samé, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cédric Féliers, and Olivier Chesneau. 2019. Mixture of Joint Nonhomogeneous Markov Chains to Cluster and Model Water Consumption Behavior Sequences. *ACM Trans. Intell. Syst. Technol.* 1, 1, Article 1 (January 2019), 22 pages. https://doi.org/10.1145/3347452

Authors' addresses: Milad Leyli-abadi; Allou Samé; Latifa Oukhellou, Paris-Est Univeristy, IFSTTAR, COSYS, GRETTIA, Marne-la-Vallée, F-77447, France, milad.leyli-abadi@ifsttar.fr, allou.same@ifsttar.fr, latifa.oukhellou@ifsttar.fr; Nicolas Cheifetz; Pierre Mandel; Cédric Féliers, Veolia Eau d'Ile-de-France, Le vermont, 28 Boulevard de Pesaro, Nanterre, F-92751, France, nicolas.cheifetz@veolia.com, pierre.mandel@veolia.com, cedric.feliers@veolia.com; Olivier Chesneau, Syndicat des Eaux d'Ile-de-France (SEDIF), 120 Boulevard Saint-Germain, Paris, F-75006, France.

2157-6904/2019/1-ART1 $15.00
https://doi.org/10.1145/3347452

1 INTRODUCTION

The issue of effective water resource management becomes increasingly important as populations grow and the climate changes. The problem is more acute for dense urban areas that involve complex water network architectures. One significant performance leverage about management and distribution of drinking water is meter readings taken over a consumption period. By analyzing the collected information, water utilities in collaboration with national authorities can establish different types of strategies, e.g., price strategies [18, 29], drought strategies [21, 28] and strategies to promote conservation [27]. The objective of these strategies is to encourage consumers to use water more responsibly. To achieve this goal, the collected consumption data must be acquired with a fine granularity. Recently, the increasing use of smart meters in the framework of smart water grids has provided water utilities with large amounts of telemetered consumption data at very low sampling rates (i.e., from fifteen seconds interval to one hour) compared to the traditional monthly sampling rate usually used for billing purposes. These finer spatial and temporal sampling resolutions make it possible to develop advanced data analytics in this domain. The data with a high longitudinal frequency are used for water end-use disaggregation [2] while the lower frequency data can be utilized for short-term consumption analysis [7, 14, 20] and other applications, such as detecting water leaks [5, 8].

One primary issue that has been addressed by several studies in the domain of water data analysis is related to clustering, in which data are reduced to an identified set of clusters to facilitate further analysis and enhance decision-making capabilities. Consumers who share common consumption behaviors are assigned to the same cluster. By partitioning the consumers into different clusters, water companies can monitor and manage the demand more efficiently. For example, one way to distinguish consumers is based on their water consumption level [1, 25]. Another possibility is to investigate the consumer consumption behavior over different time periods [23]. Then, each cluster will consist of consumers who share the same consumption habits regardless of their water consumption levels.

Various exogenous factors can also influence consumption behaviors, including environmental and water conservation attitudes [34], climatic variables [17, 37], socio-economic and demographic information [13], calendar events [36], and so on. These factors should also be considered by water utilities when analyzing consumers' behavior. Several related works have studied the influence of such covariates on water consumption using statistical approaches and feature selection methods [10, 15].

Keeping that information in mind, the main focus of this paper is on clustering of categorical sequences, which represent the evolution of consumer behavior over time. Note that in this study, clustering is performed by considering both the consumption behavior dynamics and other available exogenous factors. For this purpose, we propose a clustering approach based on a mixture of nonhomogeneous Markov models that uses contextual factors as model inputs. Aside from clustering, the proposed model can be used to perform prediction tasks. Indeed, future consumption behaviors within the identified clusters can be predicted by using the estimated parameters of each of the nonhomogeneous (input-dependent) Markov models in the mixture. We introduce several graphical tools to facilitate the interpretation of the results.

The rest of this paper is organized as follows. Section II presents the related works. Section III describes the core components of the proposed methodology for clustering and forecasting water consumption behaviors, presents the details of the estimation algorithm, and describes the analyzed real-world dataset. In Section IV, we present the experimental results using graphical and numerical tools in two separate sub-sections: the first subsection is related to the clustering results and the second subsection provides an evaluation of the forecasting performance of the proposed method.

2 RELATED WORK

Analyzing user behavior using Markov models has been widely studied in application domains such as energy, transport and web traffic analysis. In [9], the authors analyzed the navigation patterns of website users. The proposed approach partitioned users into clusters in which the users shared similar navigation paths through the site. To consider dynamic user behavior, the authors used a mixture of first-order Markov models. Each state of the Markov model corresponded to a page category that a user could visit. The authors showed that some higher order dependencies can also be captured using the mixture mechanism. Finally, the visualization results confirm the appropriateness of using Markov mixture models to cluster categorical time series. In contrast to this model, our proposed approach assumes the model to be input-dependent.

The consumption behavior of energy consumers has been analyzed in several different research papers [22, 33]. In [33], symbolic aggregate approximation was used in conjunction with the Markov models to reduce the scale of the data and to model the dynamic behavior of consumers using transition matrices. To partition consumers in different groups, a density-based clustering technique (CFSFDP) was used that takes as input the dissimilarity matrix obtained by the Kullback-Liebler distance between each pair of transition matrices. The consumption behavior of an adjacent time period is analyzed for different groups throughout an entropy measure. However, this work did not study the influence of exogenous factors on consumption behaviors. A methodology based on an encoding system with a preprocessed load shape dictionary was proposed in [22]. The load shape information of energy consumers was used to classify households with regard to extracted features such as an "entropy of shape" code which measures consumption variability. This study derived five sample programs and policy-relevant energy lifestyle segmentation strategies.

To analyze the behavior of public transport users and predict their future trajectories, the authors of [35] proposed clustering users into three categories (regular, variable and irregular) based on a spatiotemporal entropy measure and the k-means algorithm. They investigated Markov models and their hidden extensions to predict individuals' future movements within different predefined time periods throughout a day (morning, afternoon, evening, and night). The latency time (the amount of time spent at a place after exiting a metro station) and users' current locations were considered as inputs to the model. The results of a comparison indicated that the hidden Markov models were able to achieve better prediction accuracy. This model differs from ours because we used a model-based algorithm for clustering that considers the temporal dynamics of consumption behaviors. In this way, context-based clusters can be identified rather than performing grouping based on a regularity measure.

As discussed above, few papers have studied user behavior while considering the joint Markovian dynamics. In most of the developed methods consumption patterns are clustered regardless of

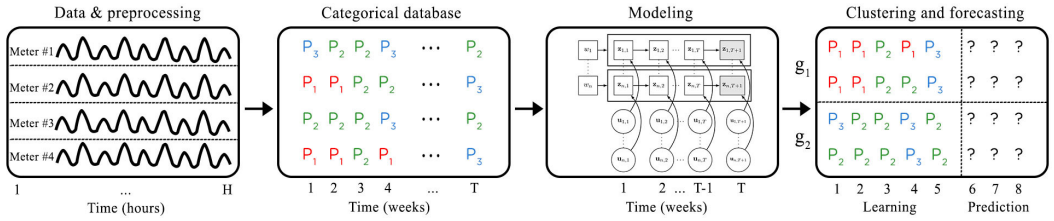


Fig. 1. Scheme of the proposed methodology, which consists of 4 steps: (1) preprocessing and normalization, (2) data discretization per week, (3) modeling consumption behavior dynamics, (4) clustering of categorical time series and forecasting of future consumption behaviors.

exogenous factors. Furthermore, to capture the consumption behavior dynamics, clustering and modeling are typically performed in two separate stages. To fill these gaps, this paper presents an integrated methodology based on a mixture of nonhomogeneous Markov models that merges clustering and modeling into a single uniform platform. In this platform, each transition between an adjacent pair of consumption behaviors (called states) is assumed to be dependent on some exogenous variables.

The contributions of this study can be summarized as follows:

- The categorical water consumption series extracted from a real-world dataset are classified using a clustering approach based on a mixture of nonhomogeneous Markov models;
- The consumption behavior dynamics related to each identified cluster are visualized using appropriate graphical representations and a quantitative metric;
- Future consumption behaviors are predicted for each cluster.

3 METHODOLOGY

The proposed methodology for clustering and forecasting the consumption behaviors consists of the different phases shown in Figure 1. The consumption profiles are normalized in a preprocessing step. The typical consumption patterns are extracted from the normalized consumption profiles, and the sequences of these patterns are used to construct a new categorical database. In this way, each meter in this categorical dataset is represented by the evolution of its users' consumption behaviors (categorical states). Next, the consumption behavior dynamics are modeled using a mixture of nonhomogeneous Markov models, leading to the estimation of a latent variable that encodes the class membership. Finally, future consumption behaviors are predicted within each cluster. These phases are described in the following subsections.

3.1 Data and preprocessing

The dataset analyzed in this paper stems from 2,000 meters located in suburban areas of Paris, France. The dataset covers a 19-month consumption period: from January 26th, 2015 to July 24th, 2016. Data collection is performed hourly using automatic meter reading (AMR) [26], and the meters mainly reflect both individual and collective housing. The data collected from the smart meters exhibit a significant number of missing values due to meter malfunctions or erroneous readings. A pretreatment stage is applied to address these missing consumption values. However, because this paper focuses on clustering and prediction of water consumption time series, we do not discuss the specific details of the preprocessing stage.

Depending on the desired forecasting time horizon (short-term, medium-term or long-term), either daily or weekly profiles can be investigated. In this paper, we investigate weekly profiles, since the aim is to be able to forecast the medium-term consumption behaviors (weekly behaviors); however, the proposed approach could also be performed by considering daily profiles. In a similar work [24], we have used the daily profiles to forecast the short-term consumption behaviors. To analyze the evolution of consumption behaviors of consumers over time, the weekly demand profiles are log-normalized. Subsequently, all the profiles are centered and reduced. The normalization has two main purposes: it scales down the consumption data for all the consumers so they present less variability, and it causes them to approach a normal distribution. These normalized data are used as inputs for the next stage.

Apart from the consumption data, this study also considers different exogenous variables such as climatic factors and calendar events. The climatic variables are temperature and precipitation level, while the calendar variables are public and school holidays.

3.2 Data discretization

To be able to analyze the massive data collected using smart meters, one solution is to reduce the data size. To accomplish this, the normalized weekly consumption curves are discretized using a functional clustering algorithm [31]. The Bayesian information criterion (BIC) [32] is used to select the number of states. As the result of clustering, we identified eight main weekly consumption behaviors or states ($K = 8$), as shown in Figure 2, where each cluster center (the solid lines) is visualized with its standard deviation (brighter colors) on the left side, and the corresponding superimposed daily patterns are shown on the right.

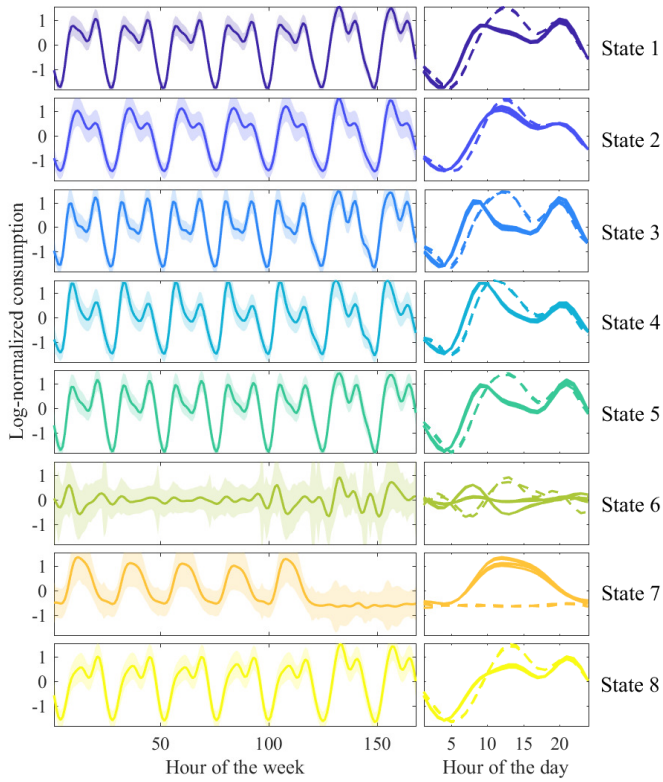


Fig. 2. Clusters centers or main weekly consumption behaviors, called states (solid lines), with their standard deviations (brighter colors) on the left and the corresponding superimposed daily profiles on the right. The plots on the right-hand side clearly reveal how consumption behavior on weekdays (solid lines) differ from those on weekends (dashed lines).

Looking at the superimposed daily patterns, the differences between consumption behaviors on weekdays (solid lines) and weekends (dashed lines) can clearly be observed. Note that for states 1–5, the morning peaks occur later in time on weekends. State 6 is characterized by a symmetrical consumption behavior on Monday and Friday with respect to Saturday and Sunday. A single peak during weekdays followed by a more consistent consumption during weekends characterizes the state 7, which can be associated with consumption behavior in a commercial activity zone. Finally, state 8 shows higher morning peaks during weekends and probably represents a consumption behavior related to holidays.

The resulted categorical dataset encoding the weekly consumption curves of each residence is shown in Figure 3. In this figure, each of the identified states (main weekly consumption behaviors) is shown by the corresponding color for each sequence.

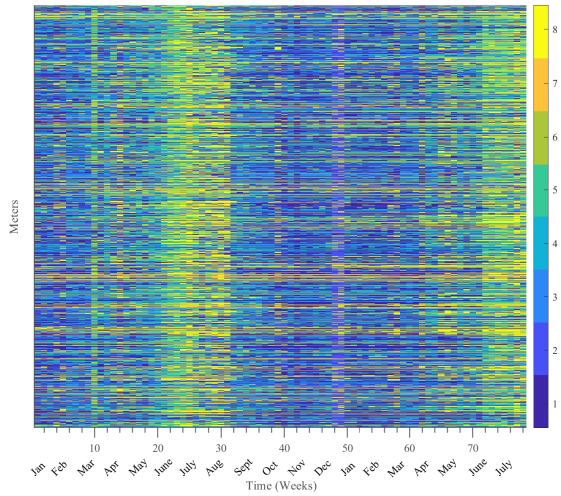


Fig. 3. Categorical dataset encoding the weekly consumption behaviors of 2,000 smart meters

NOTATIONS

$\mathbf{s}$	$(s_{it})_{i=1,\ldots,n;t=1,\ldots,T}$: dataset
s_{it}	state associated to the meter i at day t ($s_{it} \in \{1,\ldots,5\}$)
$\mathbf{s}_i$	$(s_{it})_{t=1,\ldots,T}$: consumption sequence for a meter i
$\mathbf{s}_t$	$(s_{it})_{i=1,\ldots,n}$: data of day t
w_i	label associated to the time series $\mathbf{s}_i$
$\mathbf{w}$	$(w_i)_{i=1,\ldots,n}$: set of labels associated to meters $\mathbf{s}_i$
$\mathbf{e}$	$(\mathbf{e}_{it})_{i=1,\ldots,n;t=1,\ldots,T}$: set of input variables
$\mathbf{e}_{it}$	input associated to the meter i at day t
$\boldsymbol{\phi}$	set of all the parameters of the model
$\boldsymbol{\phi}_g$	set of the parameters for the meters of cluster g
K	number of states
G	number of time series clusters

3.3 Modeling

3.3.1 Model definition. Using a set of categorical consumption time series $\mathbf{s} = \{\mathbf{s}_1,\ldots,\mathbf{s}_i,\ldots,\mathbf{s}_n\}$, where $\mathbf{s}_i$ is measured over T different weeks ($\mathbf{s}_i = \{s_{i1},\ldots,s_{iT}\}$), the proposed model assumes that each series $\mathbf{s}_i$ is issued from one of G groups. A clustering random variable $\mathbf{w} = (w_1,\ldots,w_n)$ is associated with $\mathbf{s}$ where each realization of this random variable w_i takes a value in $g \in \{1,\ldots,G\}$. In each group g, the time series are modeled by a first order nonhomogeneous Markov model from the input data or entries $\mathbf{e}$ [24], [4], [3]. This model has the following property:

$$\mathrm{P}(s_{it} \mid s_{i1},\ldots,s_{it-1},\mathbf{e}_{i1},\ldots,\mathbf{e}_{it}) = \mathrm{P}(s_{it} \mid s_{it-1},\mathbf{e}_{it}). \quad (1)$$

The current state s_{it} depends only on the previous state s_{it-1} and the input vector $\mathbf{e}_{it}$. A corresponding graphical representation of the proposed mixture model is shown in Figure 4. In this setting, we have considered that each state $\mathbf{s}_{it}$ corresponds to one of the eight identified main weekly consumption behaviors. This model is specified by the following conditional mixture density:

$$f(\mathbf{s}_i|\mathbf{e}_i;\boldsymbol{\phi}) = \sum_{g=1}^{G} p_g f_g(\mathbf{s}_i|\mathbf{e}_i;\boldsymbol{\phi}_g), \tag{2}$$

where p_g is the proportion of the cluster g that satisfies $\sum_{g=1}^{G} p_g = 1$, $f_g(.)$ is the distribution associated to the group g, and $\boldsymbol{\phi}_g$ is its set of parameters. The distribution of a time series $\mathbf{s}_i$ from a group g is given by

$$f_g(\mathbf{s}_i|\mathbf{e}_i;\boldsymbol{\phi}_g) = \mathrm{P}(s_{i1} \mid w_i = g, \mathbf{e}_{i1};\boldsymbol{\phi}_g) \prod_{t=2}^{T} \mathrm{P}(s_{it} \mid w_i = g, s_{it-1}, \mathbf{e}_{it};\boldsymbol{\phi}_g), \tag{3}$$

with

$$\mathrm{P}(s_{i1} = \ell \mid w_i = g, \mathbf{e}_{i1}) = \pi_{g\ell}(\mathbf{e}_{i1};\boldsymbol{\alpha}_g) = \frac{\exp(\boldsymbol{\alpha}_{g\ell}^{\top}\mathbf{e}_{i1})}{\sum_{\ell'=1}^{K}\exp(\boldsymbol{\alpha}_{\ell'}^{\top}\mathbf{e}_{i1})}, \tag{4}$$

$$\mathrm{P}(s_{it} = k \mid s_{it-1} = \ell, w_i = g, \mathbf{e}_{it}) = \pi_{g\ell k}(\mathbf{e}_{it};\boldsymbol{\beta}_{g\ell}) = \frac{\exp(\boldsymbol{\beta}_{g\ell k}^{\top}\mathbf{e}_{it})}{\sum_{k'=1}^{K}\exp(\boldsymbol{\beta}_{g\ell k'}^{\top}\mathbf{e}_{it})}, \tag{5}$$

where the $\pi_{g\ell}$s designate the initial input-dependent probabilities whose parameter vector is $\boldsymbol{\alpha}_g = (\boldsymbol{\alpha}_{g\ell})_{\ell=1,\ldots,K}$, and the $\pi_{g\ell k}$s designate the input-dependent transition probabilities of partition g whose set of parameters are $\boldsymbol{\beta}_{g\ell} = (\boldsymbol{\beta}_{g\ell k})_{\ell=1,\ldots,K}$. In the following, we call the proposed mixture of joint nonhomogeneous Markov models "MixJNMM."

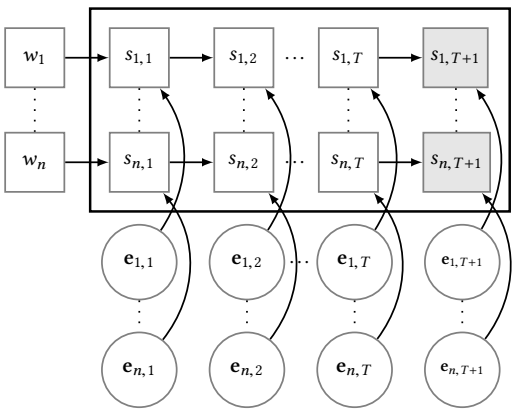


Fig. 4. A graphical representation of the proposed mixed joint nonhomogeneous Markov model (MixJNMM) with input variables

3.3.2 Parameter estimation. The optimal parameters are generally obtained by maximizing the log-likelihood

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}) &= \log \mathrm{P}(\mathbf{s}_1, \ldots, \mathbf{s}_T \mid \mathbf{e}; \boldsymbol{\phi}) = \log \prod_{i=1}^{n} \mathrm{P}(\mathbf{s}_i \mid \mathbf{e}_i; \boldsymbol{\phi}) \\ &= \sum_{i=1}^{n} \log \sum_{g=1}^{G} p_g \Big[\mathrm{P}(s_{i1} \mid w_i = g, \mathbf{e}_{i1}; \boldsymbol{\phi}_g) \prod_{t=2}^{T} \mathrm{P}(s_{it} \mid s_{it-1}, w_i = g, \mathbf{e}_{it}; \boldsymbol{\phi}_g) \Big],\end{aligned} \tag{6}$$

where $\boldsymbol{\phi} = (p_g, \boldsymbol{\alpha}_g, \boldsymbol{\beta}_g)$ is the complete set of model parameters, where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g\ell k})_{\ell k}$. The values of $\mathrm{P}(s_{i1} \mid w_i = g, \mathbf{e}_{i1}; \boldsymbol{\phi}_g)$ and $\mathrm{P}(s_{it} \mid s_{it-1}, w_i = g, \mathbf{e}_{it}; \boldsymbol{\phi}_g)$ are given by Equations (4) and (5), respectively. This criterion cannot be maximized directly. However, one can use the expectation-maximization (EM) algorithm [12] to maximize it. Considering the large global data size, we adopted the classification EM (CEM) algorithm [11] for this study because it converges faster than the EM algorithm. Using CEM, both the latent random variable $\mathbf{w}$ and the other parameters of the model are estimated by maximizing the following complete data likelihood:

$$\begin{aligned}\mathrm{P}(\mathbf{s}, \mathbf{w} \mid \mathbf{e}; \boldsymbol{\phi}) &= \mathrm{P}(\mathbf{w}) \, \mathrm{P}(\mathbf{s} \mid \mathbf{w}, \mathbf{e}; \boldsymbol{\phi}) \\ &= \prod_{i=1}^{n} \mathrm{P}(w_i) \, \mathrm{P}(\mathbf{s}_i \mid w_i, \mathbf{e}_i; \boldsymbol{\phi}_{w_i}).\end{aligned} \tag{7}$$

Developing 7, we get the following complete log-likelihood:

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\phi}, \mathbf{w}) &= \log \mathrm{P}(\mathbf{s}, \mathbf{w} \mid \mathbf{e}; \boldsymbol{\phi}) \\ &= \sum_{g=1}^{G} \Bigg[\sum_{i=1}^{n} w_{ig} \log p_g + \sum_{i=1}^{n} \sum_{\ell=1}^{K} w_{ig} s_{i1\ell} \log \pi_{g\ell}(\mathbf{e}_{i1}; \boldsymbol{\alpha}_g) \\ &+ \sum_{i=1}^{n} \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{\ell=1}^{K} w_{ig} s_{itk} s_{i(t-1)\ell} \log \pi_{g\ell k}(\mathbf{e}_{it}; \boldsymbol{\beta}_{g\ell}) \Bigg],\end{aligned} \tag{8}$$

for which we have used the following indicator variables to specify the cluster memberships and designate the transitions of the Markov model:

- $w_{ig} = 1$ if $w_i = g$ and $w_{ig} = 0$ otherwise;
- $s_{i1\ell} = 1$ if $s_{i1} = \ell$ and $s_{i1\ell} = 0$ otherwise;
- $s_{itk} = 1$ if $s_{it} = k$ and $s_{itk} = 0$ otherwise;
- $s_{itk} s_{i(t-1)\ell} = 1$ if $s_{it} = k$ and $s_{i(t-1)} = \ell$ simultaneously and $s_{itk} s_{i(t-1)\ell} = 0$ otherwise.

To promote the robustness of the parameter estimation, we add a regularization term to the likelihood function of Equation (8), which leads to the following optimization problem:

$$\arg\max_{(\boldsymbol{\phi}, \mathbf{w})} \left(\mathcal{L}_c(\boldsymbol{\phi}, \mathbf{w}) - \frac{\xi}{2} \|\boldsymbol{\phi}\|_2^2 \right), \tag{9}$$

where $\|.\|_2$ is the L_2 norm and ξ is a hyperparameter that controls the importance of the regularization term.

In the following section, we introduce the proposed CEM algorithm to estimate the parameters of the mixture model.

Proposed CEM algorithm. The CEM algorithm [11] used here is a variant of EM that incorporates a classification step between the E- and M-steps of the EM algorithm. Starting from an initial

partition $\mathbf{w}^{(0)}$, the q-th iteration of CEM for the proposed mixture approach is defined as follows:

- *E-step.* For $i = 1, \ldots, n$ and $g = 1, \ldots, G$, compute the current posterior probabilities $\tau_g^{(q)}(\mathbf{s}_i)$ that $\mathbf{s}_i$ belongs to cluster g as follows:

$$\tau_g^{(q)}(\mathbf{s}_i) = \frac{p_g^{(q)} f_g(\mathbf{s}_i; \boldsymbol{\phi}_g^{(q)})}{\sum_{g'=1}^{G} p_{g'}^{(q)} f_{g'}(\mathbf{s}_i; \boldsymbol{\phi}_{g'}^{(q)})}, \tag{10}$$

  given the current parameter estimates $\boldsymbol{\phi}^{(q)}$.

- *C-step.* The partition $\mathbf{w}^{(q+1)}$ is defined by assigning each observation $\mathbf{s}_i$ to the cluster that provides the maximum current posterior probability $\tau_g^{(q)}(\mathbf{s}_i)$, $(1 \leq g \leq G)$ as follows:

$$w_i^{(q+1)} = \arg\max_{1 \leq g \leq G} \tau_g^{(q)}(\mathbf{s}_i). \tag{11}$$

- *M-step.* Compute the maximum likelihood estimates $(p_g^{q+1}, \boldsymbol{\phi}_g^{(q+1)})$, which leads to

$$p_g^{(q+1)} = \frac{\sum_{i=1}^{n} w_{ig}^{(q)}}{n}, \qquad \forall\, g = \{1, \ldots, G\}. \tag{12}$$

Furthermore, the parameters of the model $(\boldsymbol{\alpha}_g^{(q+1)}, \boldsymbol{\beta}_{g\ell}^{(q+1)})$ can be computed by maximizing the quantity:

$$\mathcal{L}_1(\boldsymbol{\alpha}_g) + \sum_{\ell=1}^{K} \mathcal{L}_{2,\ell}(\boldsymbol{\beta}_{g\ell}), \tag{13}$$

where

$$\mathcal{L}_1(\boldsymbol{\alpha}_g^{(q+1)}) = \sum_{\ell=1}^{K} \sum_{i=1}^{n} w_{ig} s_{i1\ell} \log \pi_{g\ell}(\mathbf{e}_{i1}; \boldsymbol{\alpha}_g^{(q+1)}), \tag{14}$$

and

$$\mathcal{L}_{2,\ell}(\boldsymbol{\beta}_{g\ell}^{(q+1)}) = \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ig} s_{it-1\ell}\, s_{itk} \log \pi_{g\ell k}(\mathbf{e}_{it}; \boldsymbol{\beta}_{g\ell}^{(q+1)}). \tag{15}$$

Consequently, the problem (9) can be solved by the following $G \times (K+1)$ separate maximization problems:

$$\begin{aligned} &\arg\max_{\boldsymbol{\beta}_{g\ell}} \left[\mathcal{L}_{2,\ell}(\boldsymbol{\beta}_{g\ell}) - \frac{\xi}{2} \|\boldsymbol{\beta}_{g\ell}\|_2^2 \right], \quad \forall \;\; g, \ell \\ &\arg\max_{\boldsymbol{\alpha}_g} \left[\mathcal{L}_1(\boldsymbol{\alpha}_g) - \frac{\xi}{2} \|\boldsymbol{\alpha}_g\|_2^2 \right], \quad \forall \;\; g = 1, \ldots, G. \end{aligned} \tag{16}$$

In this paper, we opted for the Newton algorithm [30], known in this situation as the iteratively reweighted least squares (IRLS) algorithm [16]. It converges to the optimal solution at a sufficient speed.

Model selection. In the proposed method, the best model is defined as the model with the optimal value of the number of time series clusters G. To select the best model, we use the following

integrated classification likelihood (ICL) criterion [6] which is essentially the ordinary BIC [32] penalized by the entropy $-\sum_{i,g} \tau_g(\mathbf{s}_i) \log \tau_g(\mathbf{s}_i)$:

$$ICL(G) = \mathcal{L}_c(\hat{\boldsymbol{\phi}}) - \frac{\vartheta(G)}{2} \log(n \times T), \tag{17}$$

where $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\phi}$, and $\vartheta(G)$ is the number of free parameters in the model, which is given by

$$\vartheta(G) = G \times \Big[\underbrace{(K-1)m}_{\boldsymbol{\alpha}_g} + \underbrace{K \times (K-1)m}_{(\boldsymbol{\beta}_{g\ell})_{\ell=1,\ldots,K}}\Big] + \underbrace{G-1}_{(p_g)_{g=1,\ldots,G}}, \tag{18}$$

where G is the number of clusters, K is the number of states of the variables s_{it}, and m is the dimension of the input vector $\mathbf{e}_{it}$. The ICL values are computed for different numbers of parameters G. Finally, the model with the minimum ICL value is selected.

3.3.3 Forecasting. After the parameters have been estimated from the training sequence $(\mathbf{s}_1, \ldots, \mathbf{s}_T)$, the following one-step-ahead forecast permits future state prediction within each group g:

$$\begin{aligned} \widehat{s}_{iT+1} &= \arg\max_k \mathrm{P}(s_{iT+1} = k \mid s_{iT}, w_i = g, \mathbf{e}_{iT+1}) \\ &= \arg\max_k \pi_{gs_{iT}k}(\mathbf{e}_{iT+1}; \boldsymbol{\beta}_{s_{iT}}). \end{aligned} \tag{19}$$

The pseudocode in 1 summarizes the proposed MixJNMM approach.

ALGORITHM 1: MixJNMM

Data: $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$, $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$.
Result: Parameters $\widehat{\boldsymbol{\phi}}$; Partition variable $\widehat{\mathbf{w}}$.
initialization: G; $\boldsymbol{\phi}^{(0)} = (p_g^{(0)}, \boldsymbol{\alpha}_g^{(0)}, \boldsymbol{\beta}_g^{(0)})$, $\mathbf{w}^{(0)}$; fix a threshold $\epsilon > 0$; set $q \leftarrow 0$.

while $\mathcal{L}_c^{(q+1)}(\boldsymbol{\phi}) - \mathcal{L}_c^{(q)}(\boldsymbol{\phi}) > \epsilon$ **do**
  E-Step:
  forall i, g **do**
    compute $\tau_g^{(q)}(\mathbf{s}_i)$ using Eq. (10)
  end
  C-Step:
  forall i **do**
    compute $\mathbf{w}_i^{(q+1)}$ using Eq. (11)
  end
  M-Step:
  for $g = 1, \ldots, G$ **do**
    compute $p_g^{(q+1)}$ using Eq. (12)
    for $\ell = 1, \ldots, K$ **do**
      compute the parameters $\boldsymbol{\phi}_g^{(q+1)} = (\boldsymbol{\alpha}_g^{(q+1)}, \boldsymbol{\beta}_{g\ell}^{(q+1)})$ by maximizing the likelihood (13)
    end
    $q \leftarrow q + 1$
  end
  $\widehat{\boldsymbol{\phi}} = (p_g^{(q)}, \boldsymbol{\alpha}_g^{(q)}, \boldsymbol{\beta}_g^{(q)})$ and $\widehat{\mathbf{w}} = \mathbf{w}^{(q)}$.
end

4 EXPERIMENTAL RESULTS

We applied the proposed methodology to the previously discussed categorical dataset with the objectives of learning the heterogeneous structure of the consumption behaviors time series and predicting future consumption behaviors with respect to the identified structure. To this end, we used two-thirds of the consumption period (12 months of consumption during 2015) and all the available exogenous covariates to learn the mixture parameters, including the meter clustering. Then, we used the remaining data (corresponding to 7 months of consumption during 2016) to test the forecasting capabilities of the proposed method.

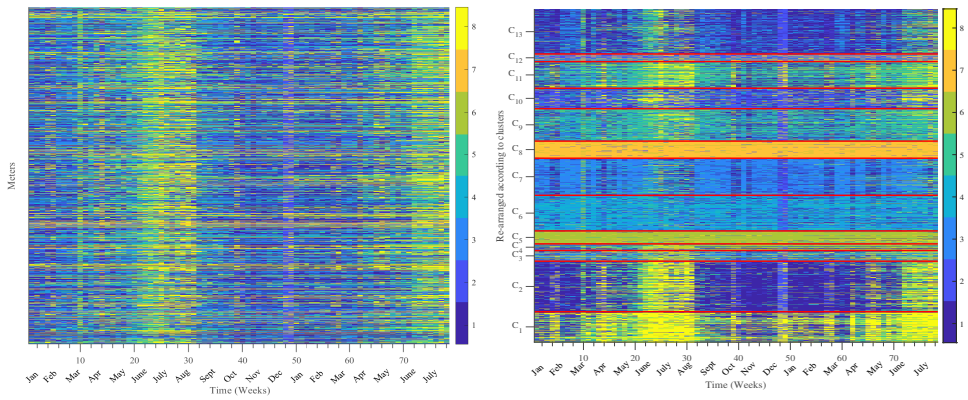
The exogenous variables used are the following climatic factors and calendar events (school holidays or summer vacations):

- T_{t-1} is the temperature (°C) at the previous timestep;
- P_{t-1} is the precipitation (in mm) at the previous timestep;
- C comprises the calendar events (school holidays or summer vacations) encoded using a binary variable at the current timestep;
- Y_{t-1} is the hourly consumption level at the previous timestep.

Although other demographic and socioeconomic information could improve the richness of the analysis, such data were not available at the time of writing. For clustering purposes, all the introduced exogenous variables are used as model inputs. However, to examine the influences of these factors on the forecasting capability of the proposed method, the input vector $\mathbf{e}$ can contain one or more of these variables. The following two sections describe the results of these stages.

4.1 Learning and clustering

The graphical representation of the proposed method (see Figure 4) shows that each meter issues from a conditional mixture density function. To be able to compute this density function for the meters in Figure 5 (a), the number of mixture components should be known beforehand. After this parameter is estimated, the learning procedure can be initiated, and the consumption behavior dynamics can be captured by considering the exogenous factors. Note that we set the hyperparameter ξ to 10^{-8}. The following sections examine these steps in more detail by providing the corresponding results.



(a) Categorical database (2,000 smart meters) (b) Clustering result

Fig. 5. Clustering process: (a) the consumption behavior sequences of 2,000 consumers and (b) the 13 clusters (separated by horizontal red lines) obtained by applying the proposed method.

4.1.1 Number of components. The number of components G of the mixture must be set beforehand, as is the case for most unsupervised clustering algorithms. To estimate G for the proposed model, we computed the ICL criterion introduced earlier in the modeling section for a varying number of components $G \in \{1, \ldots, 30\}$. The corresponding results are shown in Figure 6. A better model is obtained when the associated ICL value is minimum. In Figure 6, the minimum value of the ICL criterion is highlighted in red and corresponds to $G = 13$.

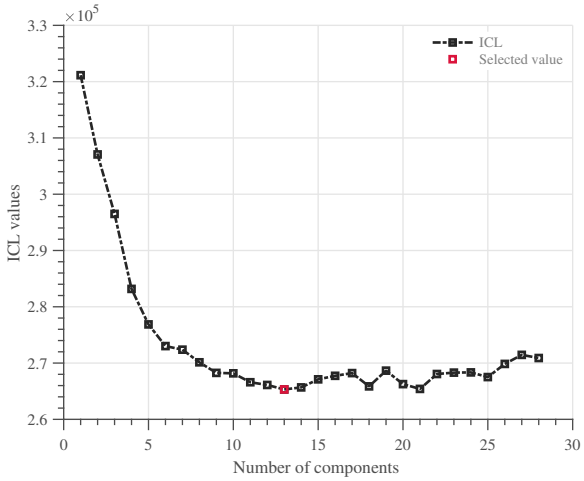


Fig. 6. ICL computed using an interval of 1–30 clusters; estimating the number of clusters.

4.1.2 Clustering result. Running the proposed mixture model with $G = 13$ allows the identification of the clusters shown in Figure 5 (b). In this figure, the meters (rows) are reorganized to reflect their cluster membership. The identified clusters are separated using horizontal red lines. Note that some clusters are made up of meters with more homogeneous consumption behaviors (e.g., clusters five and eight). To analyze the obtained clusters in more depth, the next section introduces some graphical tools and quantitative criteria.

4.1.3 Clustering sensitivity with respect to the number of states. In this section, we have conducted a sensitivity analysis of the model quality with respect to the number of states. This analysis is performed using simulated data which are constituted from three main Markovian dynamics ($G = 3$) and eight states ($K = 8$). We retrospectively varied the number of states from 2 to 20 increased by steps of 2, an operation that leads to the creation of 10 categorical datasets. Thereafter, using the proposed method (MixJNMM), we have classified the sequences in each of these datasets by fixing G to 3. Finally, we have computed the accuracy between the clustering results obtained on each dataset and the ground truth clusters of the simulated data (see Figure 7). In our case, the accuracy is estimated through the rate of the correctly classified sequences. As can be seen in Figure 7, the performances do not fluctuate significantly from $K = 8$. This can be explained by the fact that, increasing the number of states (more than 8) may result to the emergence of redundant states, which does not improve the clustering result.

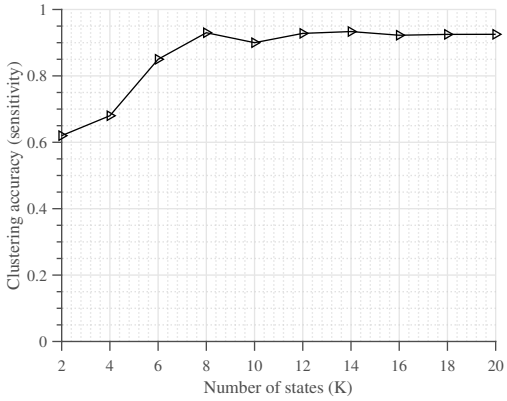


Fig. 7. Sensitivity of clustering result for different number of states K.

4.1.4 Analysis of the clustering results. One way to represent the obtained clusters is to plot the proportion of states per cluster during the consumption time period and look for significant changes in consumption behavior. In Figure 8, the state proportions are illustrated for each cluster along with the school holidays and summer vacations, which are surrounded with red and cyan dashed rectangles, respectively. The distinctions between the proportions of the consumption behaviors at different time periods is immediately apparent.

As shown in Figure 8, the state proportions change significantly during summer vacations for clusters C_1 and C_2. This change is highlighted by a significant emergence of the eighth state, which has a later morning peak compared with states 1 and 2 (both of which are prevalent in these clusters).

Taking a different case, the meters of cluster C_4 exhibit a double behavior over time—the significant emergence of state 6 after the month of June 2015. Then, this state persists until the next summer. The symmetrical behavior of state 6 could be induced by some scheduled task (e.g., a scheduled maintenance operation) undertaken starting from this date.

The more homogeneous clusters (i.e., C_5 and C_8) exhibit a stable consumption behavior over time that does not change significantly. Cluster 8 consists mainly of the state 7, which is characterized by a one-peak pattern during weekdays and more constant consumption during weekends. The meters in this cluster may be associated with a commercial activity zone.

To summarize global cluster behavior, we used the estimated Markov model transition matrices. Figure 9 represents the transition matrices associated with three of the thirteen identified clusters. These are averaged transition matrices over time. As in the case of the time-variant Markov model, the transition probabilities between states belonging to any pair of consecutive timestamps are held in a unique transition matrix. In these matrices, the transition probabilities are encoded using a range of $[0, 1]$—from white (the lowest probability) to black (the highest probability)—by $\sum_{\ell=1}^{K} \pi_{g\ell k} = 1$. The rows of these matrices indicate the states at time $t-1$, whereas the columns indicate the states at time t. In each matrix, the predominant states are surrounded by orange boxes.

By observing the general behaviors of these matrices, we can immediately perceive the difference between homogeneous and less-homogeneous clusters. In homogeneous clusters, the transition probabilities accumulate in a specific zone (e.g., cluster 8), whereas in the less-homogeneous clusters, the transition probabilities are shared between a higher number of states (e.g., clusters 2 and 9).

Considering the transition matrix of cluster 2 (Figure 9 (a)), we can see that there is a high transition probability from all the states towards state 1, which is one of the predominant states

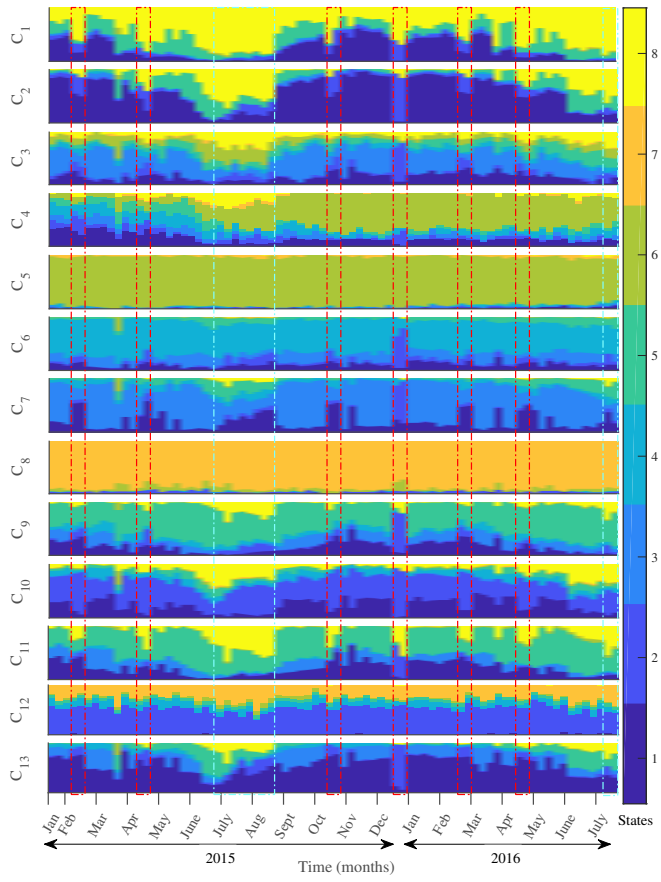
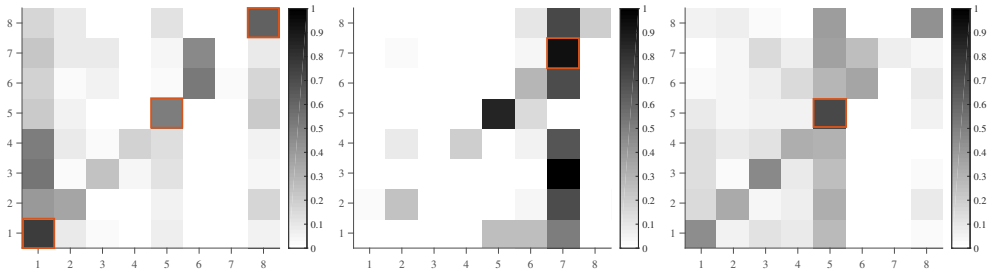


Fig. 8. State proportions over months per cluster. The states are visualized using different color codes associated with the signatures identified in Figure 2. The red dashed rectangles indicate school holidays and the cyan dashed rectangles indicate summer holidays.



(a) Cluster 2 (b) Cluster 8 (c) Cluster 9

Fig. 9. Average transition matrices associated with three of the thirteen identified clusters. Darker colors indicate higher transition probabilities. In each matrix, the predominant states are surrounded by orange boxes. The rows of the matrices indicate the states at time $t-1$, and the columns indicate the states at time t.

in this cluster. State 8, which is another predominant state in cluster 2 during summer vacations exhibits a high probability of staying in the same state. To summarize, most of the consumers associated with this cluster change their consumption behavior at the onset of summer vacations and then maintain that same behavior until the summer vacation ends.

Considering the more homogeneous cluster 8 (Figure 9 (b)), most of the transitions are towards state 7, which is the only predominant state in this cluster. The meters associated with this cluster tend not to change their consumption behavior over time, which acts as a confirmation that they belong to a spatial zone with commercial activity.

In addition to graphical representations, using a quantitative metric that measures the variability within clusters can also be helpful. For this purpose, we propose an entropy measure based on Markovian dynamics. For each observation i belonging to a cluster g, the entropy is determined as follows:

$$H(\mathbf{s}_i, g) = -\sum_t \sum_\ell \sum_k P_{k\ell}^{(g)}(\mathbf{e}_{it}) \log P_{k\ell}^{(g)}(\mathbf{e}_{it}), \tag{20}$$

where $P_{k\ell}^{(g)}(e_{it})$ indicates an input-dependent transition probability from state k to state ℓ for meter i belonging to group g. To compute the entropy at the cluster level, the mean entropy over the meters belonging to each cluster are computed as follows:

$$H(\mathbf{s}_g) = \frac{1}{n_g} \sum_{i \in g} H(\mathbf{s}_i, g), \tag{21}$$

where n_g is the number of meters in cluster g. As a result, the entropy measure is independent of the cluster size. Higher entropy values for a cluster g correspond to higher variability in the consumer consumption behaviors associated with that cluster.

The entropy distribution of the meters with respect to the clusters of Figure 5 (b) is shown in Figure 10 through boxplots. In addition, the water consumption (in cubic meters) associated with the identified clusters is superimposed over the boxplots and visualized with respect to the right vertical axis. To facilitate the interpretation, we divide the entropy distribution values into the following three regions: (i) low variability (entropy < 2), (ii) medium variability ($2 \leq$ entropy ≤ 5), and (iii) high variability (entropy > 5).

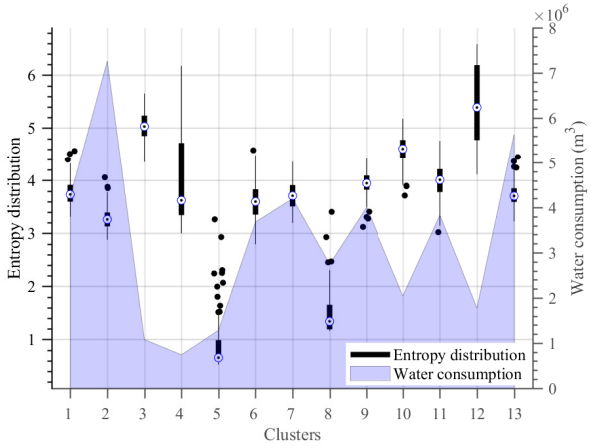


Fig. 10. Entropy distribution of meters per clusters using boxplots and mean water consumption (in cubic meters) superimposed using blue filled curves and plotted with respect to the right-hand y-axis

Looking at the entropy distribution in Figure 10, clusters 1 and 2 are characterized by medium variability in their consumption behaviors, and their water consumption is much than that of the other clusters. This confirms the fact that residential housing accounts for a large amounts of water consumption and their behaviors do not change much. In contrast, cluster 4, which encompasses only a few meters, exhibits medium-to-high variability in consumption behavior and has the lowest consumption volume. This follows the interpretation given previously when discussing Figure 8. We can observe that a very low entropy distribution exists for the homogeneous clusters 5 and 8, which consist mainly of one state; however, cluster 8 includes a higher number of meters and, consequently, a higher consumption volume.

Another interesting pattern that can be observed regarding the entropy is the relationship between cluster size, water consumption volume per cluster, and the variability inside the clusters. In most cases, high variability is associated with the clusters that include fewer meters, yet at the same time, present a lower consumption volume. These clusters present irregular behaviors over time and can be the subject of further analysis.

4.1.5 Analysis of the influence of the context variables. To conduct this analysis, we relied on the estimated coefficients of the model to provide more details concerning the influence of the exogenous factors on consumption behaviors. To this end, we have extracted the estimated coefficients $\boldsymbol{\beta}$ corresponding to climatic variables (i.e., temperature and precipitation) for the most significant transitions of two different clusters (i.e., C_2 and C_9). First, let us recall that, within a cluster g, the transition probability from a state ℓ to a state k is defined by

$$\mathrm{P}(s_{it} = k \mid s_{it-1} = \ell, w_i = g, \mathbf{e}_{it}) = \pi_{g\ell k}(\mathbf{e}_{it}; \boldsymbol{\beta}_{g\ell}) = \frac{\exp\left(\boldsymbol{\beta}_{g\ell k}^{\top} \mathbf{e}_{it}\right)}{\sum_{k'=1}^{K} \exp\left(\boldsymbol{\beta}_{g\ell k'}^{\top} \mathbf{e}_{it}\right)}.$$

The above mentioned coefficients for clusters 2 and 9 are shown respectively in the tables 1 and 2. In these tables, each value corresponds to the coefficient of the above logistic regression model, for a climatic variable, given transitions from the states located in the rows to the states located in the columns. The most significant values are highlighted in bold face in these tables. Cluster 2 is dominated by the state 8 during the summer period and by state 1 during the remaining period (see figure 5 (b)). Regarding cluster 9, it is constituted mainly by states 1, 5 and 8.

Looking at Table 1 (a), we can notice a high positive coefficient value for the transitions from state 1 to states 5 and 8. An increase in temperature therefore results in an increase of such transitions (occurring mainly when entering the summer period). Looking at the Table 1 (b), regarding the precipitation variable coefficients, we can notice a high positive value for the transition from the state 8 to the state 1, following which an increase in rainfall may lead to an increase of such transitions (occurring mainly when entering the autumn period). Additionally, we can notice the high negative values for transitions from states 1 and 8 to state 2, according to which a low precipitation rate may lead to an increase of such transitions.

Table 2 represents the transition coefficients of cluster 9 corresponding to the above-mentioned climatic variables. Looking at the Table 2 (a), we can notice a high positive value for the transition from the state 8 to itself and a high negative value for the transition from the state 5 to the state 1. This would signify that high temperatures lead to an increasing number of transitions from state 8 to itself and low temperatures lead to an increasing number of transitions from state 5 to state 1. Looking at Table 2 (b), we can notice a high positive coefficient value for the transition form state 8 to state 5. It means that a high precipitation rate leads to an increasing number of such transitions for this cluster.

Table 1. Estimated coefficients of the most significant transition probabilities of cluster 2 for temperature and precipitation factors

(a) Temperature coefficients

s_{t-1} \ s_t	1	2	5	8
1	0.17	0.19	**0.30**	**0.28**
8	-0.12	-0.06	-0.01	0.12

(b) Precipitation coefficients

s_{t-1} \ s_t	1	2	5	8
1	2.13	**-3.63**	0.68	0.41
8	**3.90**	**-4.77**	2.30	-1.37

Table 2. Estimated coefficients of the most significant transition probabilities of cluster 9 for temperature and precipitation factors

(a) Temperature coefficients

s_{t-1} \ s_t	1	5	8
1	-0.11	-0.03	-0.01
5	**-0.17**	-0.06	0.03
8	-0.03	0.08	**0.14**

(b) Precipitation coefficients

s_{t-1} \ s_t	1	5	8
1	0.46	1.75	0.44
5	1.61	2.42	-1.49
8	0.46	**3.89**	-0.69

4.2 Forecasting

Using the mixture parameters and the input-dependent Markov model transition matrices estimated during the learning process, future consumption behaviors can be predicted within each cluster. As mentioned earlier, we used the first 7 months of data from 2016 to test the performance of the proposed method. Here, we provide the details of the experimental setup used to evaluate the performance of the proposed method in terms of the prediction error.

4.2.1 Evaluated methods. The performance of the proposed method in terms of forecasting accuracy is evaluated with respect to four other methods:

- The homogeneous Markov model (MM) is a time-independent model and assumes that a single transition matrix summarizes the evolution of consumption behavior dynamics for all the categorical sequences;
- The mixture of homogeneous Markov models (MixMM) fits a homogeneous Markov model within each cluster to model the dynamics of the consumption behaviors;
- The joint nonhomogeneous Markov model (JNMM), unlike the MM, considers the temporal regularity of states;
- The K-means+JNMM is a combination of the k-means algorithm and the nonhomogeneous Markov model. The initial time series (with an hourly frequency) are clustered in the first step by the k-means algorithm; then, the categorical series within each cluster are predicted using the nonhomogeneous Markov model.

4.2.2 Evaluation criteria. To compare the forecasting results, we compute different criteria for each configuration as follows:

- The adjusted Rand index (ARI) [19] is a measure of agreement between two clustering results and is a multi-class criterion;
- Accuracy is the percentage of correctly predicted daily consumption behaviors;
- Precision is defined for each category by the rate of correctly predicted instances among the instances affected by this category;
- Recall is defined for each category by the rate of correctly predicted instances among the instances of that category;
- the F-measure represents the harmonic mean of precision and recall.

4.2.3 Comparisons. Table 3 summarizes the comparison results between the aforementioned methods when using different combinations of input variables. Note that no input variable exists for the methods based on the homogeneous Markov model (i.e., MM and MixMM) because it is inherently time-invariant. The evaluation results obtained by each criterion appear as bars inside each associated table cell. All the evaluated criteria can take values within $[0, 1]$, where 1 indicates the best performance. The bar lengths are designed to be proportional to the values of the metrics and to facilitate the interpretation of the results. The best performances are highlighted with red bars.

Table 3. Comparison of the models in terms of the supervised metrics during the 26 test weeks: MM: homogeneous Markov model, MixMM: mixture of homogeneous Markov models, JNMM: joint nonhomogeneous Markov model, k-means+JNMM: joint nonhomogeneous Markov model within the clusters identified by the k-means algorithm, MixJNMM: mixture of nonhomogeneous Markov models, Y: meter water consumption, T: temperature, P: precipitation, C: calendar information

Models	Inputs ($\mathbf{e}_{it}$)	ARI	Accuracy	Recall	Precision	F-Measure
MM		0.36	0.65	0.67	0.67	0.67
MixMM		0.37	0.65	0.68	0.68	0.68
JNMM	(T_{t-1}, P_{t-1})	0.57	0.66	0.67	0.67	0.67
	(Y_{t-1})	0.59	0.71	0.72	0.73	0.72
	$(Y_{t-1}, T_{t-1}, P_{t-1})$	0.62	0.71	0.72	0.73	0.73
	(Y_{t-1}, C)	0.61	0.73	0.72	0.72	0.72
	$(Y_{t-1}, C, T_{t-1}, P_{t-1})$	0.63	0.74	0.73	0.74	0.73
K-means + JNMM	(T_{t-1}, P_{t-1})	0.58	0.69	0.67	0.69	0.68
	(Y_{t-1})	0.61	0.74	0.74	0.72	0.73
	$(Y_{t-1}, T_{t-1}, P_{t-1})$	0.63	0.75	0.75	0.76	0.75
	(Y_{t-1}, C)	0.62	0.74	0.75	0.74	0.74
	$(Y_{t-1}, C, T_{t-1}, P_{t-1})$	0.65	0.76	0.76	0.75	0.76
MixJNMM	(T_{t-1}, P_{t-1})	0.59	0.71	0.73	0.71	0.72
	(Y_{t-1})	0.64	0.79	0.77	0.78	0.77
	$(Y_{t-1}, T_{t-1}, P_{t-1})$	0.67	0.80	0.81	0.79	0.80
	(Y_{t-1}, C)	0.64	0.79	0.78	0.76	0.77
	$(Y_{t-1}, C, T_{t-1}, P_{t-1})$	0.70	0.82	0.81	0.80	0.80

By looking at this table, we can see that the methods depending on the homogeneous Markov chains (MM and MixMM) achieve low forecasting precision levels. In contrast, the models that take exogenous factors as input (JNMM, K-means+JNMM and MixJNMM) obtain their best results when the input vector $\mathbf{e}$ includes all the available covariates (temperature, precipitation, calendar events

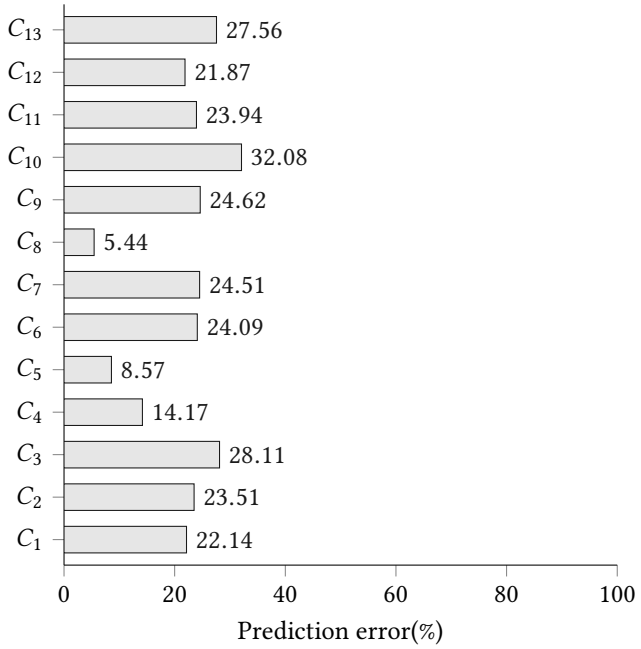


Fig. 11. Prediction error with respect to the clusters using the proposed MixJNMM and the following inputs: temperature, precipitation and calendar events

and consumption level). Figure 8 illustrated how calendar events can affect consumer consumption behavior over time.

The proposed MixJNMM method obtains the best prediction results. The proposed method may give better results due to the presence of heterogeneous latent structure among the different categorical time series. The mixture property of the proposed method allows for the proper identification of this structure. Its Markovian property provides us with the ability to capture the joint dynamics (state-state and state-input) of the consumption behaviors.

Figure 11 provides greater detail concerning the forecasting results when using the proposed method. In this figure, the prediction errors are measured for each identified cluster and are expressed as percentages. We can see that the prediction error changes significantly between clusters. For the clusters that exhibit homogeneous consumption behaviors over time (i.e., clusters 5 and 8), the prediction error remains very low. In contrast, as the consumption behavior variability increases inside the clusters, the prediction error also increases.

5 DISCUSSION AND CONCLUSION

In this paper, we propose a novel methodology based on a mixture of Markov models for analyzing consumer behavior. Using the methodology proposed in this study, categorical time series representing sequences of consumption behaviors were classified into homogeneous groups while considering the joint Markovian dynamics. Subsequently, we predicted the future consumption behaviors of each identified cluster using a set of input variables and cluster-specific estimated parameters. The numerical and graphical evaluations on a real-world dataset demonstrated the effectiveness of the proposed methodology for clustering and forecasting of water consumption time series. In this article, we have analyzed weekly consumption behaviors, however, the proposed

method could potentially be extended to treat daily behaviors and can also be applied to other application domains involving categorical time series.

The proposed method allows clustering of water consumers and forecasting of their future consumption habits. This method can help water utilities in different ways to better manage the water network. By clustering the consumers in different groups, inside which they share similar consumption habits, it can accelerate the decision making process and forecasting the future consumption habits could be necessary in some critical conditions such as a drought period for better distribution of this valuable resource. It provides also water utilities with the ability of targeting a group of consumers based on their consumption behavior dynamics.

As an insight for future works, in order to improve the forecasting accuracy of the proposed method, the transition dynamics can be learned on a set of homogeneous segments in each cluster. It requires the model modification, following which, each segment should be composed of adjacent time instants in order to respect the time dependence constraint of the proposed method. Additionally, spatial localization of smart meters may be used in future works to expand the analysis of consumption behaviors and to assign spatial implications to clusters. The proposed method may also be adapted to detect consumption behavior changes and conduct change analyses."

ACKNOWLEDGMENTS

This work was conducted within the framework of the project "Analyse de données massives relevées à distance sur les compteurs d'eau" involving Veolia Eau d'Ile de France, the Syndicat des Eaux d'Ile de France (SEDIF) and the French institute of science and technology for transport, development and networks (IFSTTAR).

REFERENCES

[1] K Aksela and M Aksela. 2010. Demand estimation with automated meter reading in a distribution network. *Journal of Water Resources Planning and Management* 137, 5 (2010), 456–467.

[2] Cara D Beal, Rodney A Stewart, and Kelly Fielding. 2013. A novel mixed method smart metering approach to reconciling differences between perceived and actual residential end use water consumption. *Journal of Cleaner Production* 60 (2013), 116–128.

[3] Yoshua Bengio. 1999. Markovian models for sequential data. *Neural computing surveys* 2, 199 (1999), 129–162.

[4] Yoshua Bengio and Paolo Frasconi. 1996. Input-output HMMs for sequence processing. *IEEE Transactions on Neural Networks* 7, 5 (1996), 1231–1249.

[5] Andrew Berglund, Venkata Siva Areti, and G Mahinthakumar. 2017. Successive linear approximation methods for leak detection in water distribution systems. *Journal of Water Resources Planning and Management* 143, 8 (2017), 04017042.

[6] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* 22, 7 (2000), 719–725.

[7] John Bougadis, Kaz Adamowski, and Roman Diduch. 2005. Short-term municipal water demand forecasting. *Hydrological Processes: An International Journal* 19, 1 (2005), 137–148.

[8] Tracy C Britton, Rodney A Stewart, and Kelvin R O'Halloran. 2013. Smart metering: enabler for rapid and effective post meter leakage identification and water loss management. *Journal of Cleaner Production* 54 (2013), 166–176.

[9] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. 2003. Model-based clustering and visualization of navigation patterns on a web site. *Data mining and knowledge discovery* 7, 4 (2003), 399–424.

[10] Rachel Cardell-Oliver. 2013. Discovering water use activities for smart metering. In *Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on*. IEEE, IEEE, 171–176.

[11] Gilles Celeux and Gérard Govaert. 1992. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis* 14, 3 (1992), 315–332.

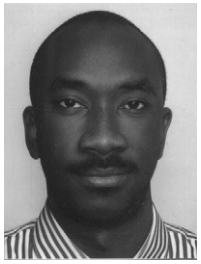
[12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

[13] Elena Domene and David Saurí. 2006. Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona. *Urban Studies* 43, 9 (2006), 1605–1623.

[14] Francesca Gagliardi, Stefano Alvisi, Zoran Kapelan, and Marco Franchini. 2017. A probabilistic short-term water demand forecasting model based on the Markov Chain. *Water* 9, 7 (2017), 507.

[15] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.
[16] Paul W Holland and Roy E Welsch. 1977. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods* 6, 9 (1977), 813–827.
[17] Lily House-Peters, Bethany Pratt, and Heejun Chang. 2010. Effects of Urban Spatial Structure, Sociodemographics, and Climate on Residential Water Consumption in Hillsboro, Oregon 1. *JAWRA Journal of the American Water Resources Association* 46, 3 (2010), 461–472.
[18] Charles W Howe and Frank Pierce Linaweaver. 1967. The impact of price on residential water demand and its relation to system design and price structure. *Water Resources Research* 3, 1 (1967), 13–32.
[19] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2, 1 (1985), 193–218.
[20] Ashu Jain, Ashish Kumar Varshney, and Umesh Chandra Joshi. 2001. Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks. *Water resources management* 15, 5 (2001), 299–321.
[21] Douglas S Kenney, Christopher Goemans, Roberta Klein, Jessica Lowrey, and Kevin Reidy. 2008. Residential water demand management: lessons from Aurora, Colorado1. *JAWRA Journal of the American Water Resources Association* 44, 1 (2008), 192–207.
[22] Jungsuk Kwac, June Flora, and Ram Rajagopal. 2014. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid* 5, 1 (2014), 420–430.
[23] Wouter Labeeuw and Geert Deconinck. 2013. Residential electrical load model based on mixture model clustering and Markov models. *IEEE Transactions on Industrial Informatics* 9, 3 (2013), 1561–1569.
[24] Milad Leyli-Abadi, Allou Same, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cédric Feliers, and Olivier Chesneau. 2017. Predictive Classification of Water Consumption Time Series using Non-homogeneous Markov Models. In *IEEE DSAA 2017, International Conference on Data Science and Advanced Analytics*. IEEE, 8p.
[25] SA McKenna, F Fusco, and BJ Eck. 2014. Water demand pattern classification from smart meter data. *Procedia Engineering* 70 (2014), 1121–1130.
[26] Kimbel A Nap, Lance A Ehrke, and Donn R Dresselhuys. 2001. Automatic meter reading data communication system. US Patent 6,246,677.
[27] Sheila M Olmstead and Robert N Stavins. 2009. Comparing price and nonprice approaches to urban water conservation. *Water Resources Research* 45, 4 (2009).
[28] Sandra L Postel. 2000. Entering an era of water scarcity: the challenges ahead. *Ecological applications* 10, 4 (2000), 941–948.
[29] Bill Randolph and Patrick Troy. 2008. Attitudes to conservation and water consumption. *environmental science & policy* 11, 5 (2008), 441–455.
[30] Kees Roos, Tamás Terlaky, and Jean-Philippe Vial. 1998. Theory and algorithms for linear optimization - an interior point approach. In *Wiley-Interscience series in discrete mathematics and optimization*.
[31] Allou Samé, Zineb Noumir, Nicolas Cheifetz, Anne-Claire Sandraz, and Cédric Féliers. 2016. Décomposition et classification de données fonctionnelles pour l'analyse de la consommation d'eau. In *Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2016), Atelier Clustering et Co-clustering (CluCo 2016)*. 11p.
[32] G. Schwarz et al. 1978. Estimating the dimension of a model. *The annals of statistics* 6, 2 (1978), 461–464.
[33] Yi Wang, Qixin Chen, Chongqing Kang, and Qing Xia. 2016. Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE transactions on smart grid* 7, 5 (2016), 2437–2447.
[34] Rachelle M Willis, Rodney A Stewart, Kriengsak Panuwatwanich, Philip R Williams, and Anna L Hollingsworth. 2011. Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. *Journal of environmental management* 92, 8 (2011), 1996–2009.
[35] Chao Yang, Fenfan Yan, and Satish V Ukkusuri. 2018. Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system. *Transportmetrica A: Transport Science* 14, 7 (2018), 576–597.
[36] SL Zhou, TA McMahon, A Walton, and J Lewis. 2002. Forecasting operational demand for an urban water supply zone. *Journal of hydrology* 259, 1-4 (2002), 189–202.
[37] Shuang Lin Zhou, Thomas Aquinas McMahon, Allan Walton, and Jane Lewis. 2000. Forecasting daily urban water demand: a case study of Melbourne. *Journal of hydrology* 236, 3-4 (2000), 153–164.

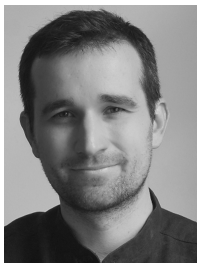
Milad Leyli-abadi received an M.S. from the Department of Machine Learning for Data Science, Paris Descartes University, Paris, France, in 2016. He is currently pursuing a Ph.D. in machine learning at the French Institute of Science and Technology for Transport, Development and Networks (Ifsttar) and is attached to the university of Paris-Est Créteil (UPEC). His research interests include time series modeling, forecasting models, machine learning and big data analytics.



Allou Samé received the Ph.D. degree in the field of real-time data clustering from Compiègne University of Technology in 2004. He received the Habilitation à Diriger des Recherches degree in temporal and functional data modeling from Paris-Est University in 2014. Since 2006, he has been a Researcher with the French Institute of Science and Technology for Transport, Development and Networks. He has been managing the Data and Mobility Group, COSYS/GRETTIA Laboratory since 2016. He has authored or co-authored over 60 papers in scientific journals and conference proceedings. His research interests include unsupervised statistical learning, pattern recognition, temporal data analysis, and their application to the diagnosis of transportation systems and the analysis of water and energy networks.

Latifa Oukhellou received a Ph.D. from Paris-Sud University in 1997 and Habilitation à diriger des Recherches from Paris-Est University in 2010. She is currently a Researcher Director at the French Institute of Science and Technology for Transport, Development and Networks (Ifsttar) in France and has been Assistant Professor at University of Paris-Est Créteil (UPEC). Her research interests concern Data Analytics, machine learning and information fusion applied to diagnosis problems as well as to spatio-temporal data mining for identifying driving behavior, analyzing urban mobility or monitoring energy and water smart grids. She has published more than 80 papers in international scientific journals and conference proceedings and she is involved in several research projects in the field of intelligent transportation systems or urban computing for smart cities.

Nicolas Cheifetz received the M.Sc. degree in AI from University Pierre and Marie Curie, Paris, in 2009 and his Ph.D. degree in data analysis from Paris-Est University, France, in 2013. Since, he works at Veolia Water as a data scientist for solving problems related to the production and distribution of drinking water. His research interests include time series analysis, graph processing and unsupervised learning.



Pierre Mandel received an M. Sc. degree in materials science from École des Mines de Saint-Etienne (France) and an M. Sc. degree in chemical engineering from Technische Universität Berlin (Germany). He holds a PhD in chemical engineering from Université de Rennes I and has been working for ten years on various research projects for Veolia. His research interests include data clustering, meta-heuristics algorithms and multiobjective optimization.