Modélisation dynamique, classification et détection de changement dans les panels catégoriels issus d'un réseau d'eau intelligent

Milad Leyli Abadi

UNIVERSITÉ PARIS-EST
ÉCOLE DOCTORALE MSTIC
MATHÉMATIQUES ET SCIENCES ET TECHNOLOGIES
DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE

POUR OBTENIR LE GRADE DE

DOCTEUR EN SCIENCES

DÉLIVRÉ PAR

L'UNIVERSITÉ PARIS-EST

SPÉCIALITÉ : INFORMATIQUE

PRÉSENTÉE ET SOUTENUE PAR

MILAD LEYLI ABADI

MODÉLISATION DYNAMIQUE, CLASSIFICATION ET DÉTECTION DE CHANGEMENT DANS LES PANELS CATÉGORIELS ISSUS D'UN RÉSEAU D'EAU INTELLIGENT

soutenue publiquement le 12 décembre 2019 devant le jury composé de :

M. GREGORY NUEL *Directeur de recherche CNRS, INSMI*	RAPPORTEUR
M. CHRISTOPHE AMBROISE *Professeur, Université d'Évry Val d'Essonne*	RAPPORTEUR
M. FABRICE ROSSI *Professeur, Université de Paris-Dauphine*	EXAMINATEUR
M. PIERRE MANDEL *Chef de projet, Veolia Eau d'Île-de-France*	EXAMINATEUR
MME. LATIFA OUKHELLOU *Directrice de recherche, IFSTTAR*	CO-DIRECTRICE DE THÈSE
M. ALLOU SAMÉ *Chargé de recherche HDR, IFSTTAR*	DIRECTEUR DE THÈSE

Modélisation Dynamique, Classification et Détection de Changement Dans les Panels Catégoriels Issus d'un Réseau d'Eau Intelligent

Manuscrit mis en page avec LaTeX 2ε à partir d'une version personnalisée du template proposé par Dorian Depriester.

Remerciements

Je remercie tout particulièrement M. Allou SAMÉ, mon directeur de thèse pour tout le soutien scientifique qu'il m'a apporté durant cette thèse, ses précieux conseils, ses relectures, sa patience et sa disponibilité qui m'ont permis d'aboutir ce travail. Je remercie également Mme. Latifa OUKHELLOU, mon encadrante de thèse pour sa présence dans les moments importants, ses relectures et ses conseils. Leur engagement et leur regard bienveillant à mon égard m'ont permis d'enrichir mes connaissances, d'acquérir des méthodes de travail et de progresser dans la rédaction de mes travaux durant ces trois années de thèse.

Je tiens à remercier tous les membres du jury pour m'avoir fait l'honneur de participer à l'évaluation de mes travaux de thèse. Je remercie notamment M. Christophe AMBROISE et M. Gregory NUEL, qui ont accepté de rapporter ces travaux. Je remercie également M. Fabrice ROSSI, examinateur de thèse, pour l'attention et l'intérêt qu'il a porté à mes travaux.

Je remercie également l'ensemble de l'équipe de Veolia (M. Pierre MANDEL, M. Nicolas CHEIFETZ et M. Cédric FÉLIERS) et celle du SEDIF (M. Olivier CHESNEAU) pour nos nombreux échanges, leur écoute, leur soutien et leurs encouragements à poursuivre et à approfondir ces recherches. Les nombreux points d'avancement que nous avons faits ensemble m'ont permis de synthétiser et de présenter régulièrement mes travaux ce qui m'a permis de mener à bien cette thèse tout en ayant un aperçu de ce que peut être la recherche en entreprise.

Je remercie l'ensemble du personnel du GRETTIA que j'ai pu côtoyer tout au long de cette thèse ainsi que les thésards du laboratoire avec lesquels j'ai passé des moments agréables.

Je remercie chaleureusement mes parents pour leur soutien et encouragements durant ces trois années de thèse. Enfin, je remercie Julie pour m'avoir soutenu, pour sa patience et ses encouragements quotidiens.

Résumé

De nos jours, on observe une préoccupation croissante suscitée par les problèmes environnementaux et ceux liés à la gestion des ressources comme l'eau et l'électricité. Dans le cadre d'une collaboration avec Veolia Eau d'Île-de-France et le syndicat des eaux d'Île-de-France, cette thèse se focalise dans un premier temps sur la classification des consommateurs d'eau ayant une dynamique d'habitudes de consommation similaire dans le temps. Dans chaque classe, cette dynamique dépend d'un nombre de facteurs exogènes. Pour modéliser cette densité jointe, nous utilisons un modèle de mélange où chaque composante est un modèle de Markov non homogène. Une fois les paramètres de ce modèle estimés, les futures habitudes de consommation dans chaque classe peuvent être prédites. Dans un second temps, le problème de la détection de changements structurels, communs à un ensemble de consommateurs est également étudié. Pour ce faire, les tests séquentiels d'hypothèses du rapport de vraisemblance sont utilisés. Ces derniers sont fondés sur des modèles de Markov non homogènes pour pouvoir modéliser le comportement dynamique des consommateurs. Un seuil adaptatif est également estimé en utilisant les simulations de type Monte Carlo. Cela permet d'adapter le seuil à différents types de changement et de réduire le taux de fausses alarmes. Les résultats de classification et de détection de changements obtenus sur une base de données réelle issue d'un réseau d'eau se sont révélés pertinents et efficaces. Finalement, une analyse de l'influence des variables exogènes en utilisant les paramètres estimés des modèles proposés permet d'enrichir les interprétations.

Mots clés: réseau d'eau intelligent, compteur communicant, consommation d'eau, séquences catégorielles, modèles de Markov non homogènes, modèle de mélange, algorithme EM, prévision, détection de points de changement, tests d'hypothèses séquentiels

Abstract

Dynamic modelling, clustering and change detection in categorical panel data issued from smart water grids

Nowadays, we observe a growing concern raised by the environmental issues and those related to management of the resources as electricity and water. As part of a collaborative project with *Veolia Eau d'Île-de-France* and *le syndicat des eaux d'Île-de-France*, this PhD research addresses initially the clustering of water consumers based on their consumption behavior dynamics over time. These dynamics, in each cluster, depend on a number of exogenous factors. To model this joint density, non-homogeneous Markov models are investigated as the components of a mixture model. Hence, the estimation of the parameters in each cluster allows to predict the future consumption behaviors independently. Afterwards, the problem of online structural change detection in a set of consumption behavior sequences is addressed. To this end, a sequential hypothesis testing of generalized likelihood ratio, based on a non-homogeneous Markov model is proposed. An adaptive threshold is also used which can be adjusted throughout the various types of changes and may reduce the number of false alarms. The results on a real dataset which is issued from a water network allow to highlight the effectiveness of the proposed methods both in terms of clustering and change detection. Finally, the analysis of the estimated parameters of both models allows to study the influence of exogenous factors on clustering and detected changes.

Keywords: smart water grids, smart meters, water consumption, categorical sequences, behavior dynamics, non-homogeneous Markov models, mixture models, EM algorithm, forecasting, change-point detection, sequential hypothesis testing

Table des matières

Liste des figures

Liste des tableaux

Notation

Notation générale

H	Longueur des séries temporelles initiales (nombre d'heures)
$\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$	Ensemble de n séries temporelles initiales
$\mathbf{y}_i = (y_1, \ldots, y_{\mathrm{H}})$	Une série temporelle de taille H
S	Nombre de profils saisonnier (journalier ou hebdomadaire)
$\mathbf{x} = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i\mathrm{S}})$	Ensemble de S profils saisonniers pour une séquence i
$\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$	Ensemble des covariables associées à des observations $\mathbf{x}$
$u_{it} \in \mathbb{R}^m$	Vecteur des variables exogènes observé au temps t et de dimension m
$f(\cdot)$	Fonction de masse (probabilité) d'un mélange de lois
$f_k(\cdot)$	Fonction de masse au sein d'une composante du mélange de lois k
$f_k(\mathbf{y}_i; \cdot)$	Fonction de masse pour une observation i au sein d'une classe k
θ	Vecteur des paramètres
θ_k	Vecteur des paramètres associé à composante k du mélange de lois
$\hat{\theta}$	Estimation de θ
p_k	Probabilité a priori associé à composante k du mélange de lois
$\mathcal{N}(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$	Loi normale/gaussienne de moyenne $\boldsymbol{\mu}_k$ et de matrice de covariance $\boldsymbol{\Sigma}_k$
$\mathrm{P}(x; \theta)$	Densité de probabilité pour une variable x et de paramètre θ
$\mathbb{E}(x; \theta)$	Espérance mathématique de x associée à une distribution paramétrée par θ
$\mathcal{L}(\theta)$	Log-vraisemblance de θ
$\mathcal{CL}(\theta)$	Log-vraisemblance complétée de θ

Modèle de Markov

K	Nombre d'états pour un modèle de Markov
$\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$	Ensemble de n séquences catégorielles
$\mathbf{z}_i = (z_{i1}, \ldots, z_{i\mathrm{T}})$	Une séquence catégorielle ou séquence d'états de taille T
G	nombre de groupes des séquences dans le mélange de modèles de Markov
$\mathbf{w} = (w_1, \ldots, w_n)$	Variable latente indiquant l'appartenance de chaque séquence à un groupe
$(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_{g\ell})$	Paramètres de probabilités initiales et celles de transition pour une composante du mélange de modèles de Markov
$\pi_{\ell k}(\cdot)$	Probabilités de transition de l'état ℓ à l'état k

Détection de changement

H_0	Hypothèse zéro dans les tests statistiques (pas de changement)
H_1	Hypothèse alternative dans les test statistiques (changement)
P_{θ_0}	Densité de probabilité avant le point de changement
P_{θ_1}	Densité de probabilité après le point de changement
Λ	Statistique de test (rapport de vraisemblance)
h	Seuil de détection
$\mathcal{Q}_{1-p}$	Quantile d'ordre $1-p$ d'une loi de probabilité

Chapitre 1

Introduction générale

Sommaire

1.1 Contexte industriel

De nos jours, on observe une préoccupation croissante suscitée par les problèmes environnementaux et ceux liés à la gestion des ressources comme l'eau et l'électricité (DAY et CONWAY, 2009; HUNAIDI et collab., 2005). Les pays sont ainsi amenés à prendre des mesures visant à une meilleure rationalisation de celles-ci, dans une optique de développement durable. Le concept nouveau des réseaux intelligents appelés *Smart Grids* offre la possibilité, à travers des technologies avancées de l'information et de la communication, de mieux gérer ces ressources en termes de disponibilité et de fiabilité, tout en incluant l'aspect économique. Si le relevé des consommations se faisait traditionnellement une fois par mois, les compteurs intelligents appelés *Smart Meters* autorisent désormais une lecture horaire voire quotidienne des consommation. Ces possibilités accrues de recueil de données permettent de surveiller de manière plus fine l'usage de l'eau ou de l'électricité, de manière à être en mesure de détecter plus rapidement des anomalies dans les réseaux (fuites d'eau, gaspillage).

Ces travaux de thèse s'inscrivent dans le cadre d'une collaboration entre Veolia Eau D'Île-de-France, le Syndicat des Eaux D'île-de-France (SEDIF) et le laboratoire GRETTIA de l'Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR). Cette collaboration porte sur l'analyse de données de consommation d'eau potable issues de compteurs du réseau SEDIF. Les travaux effectués dans cette thèse ont fait l'objet de développement d'algorithmes dont une interface graphique permettant de regrouper les consommateurs d'eau.

Dans la suite, un réseau d'eau intelligent et ses composantes sont d'abord présentés de manière générale. Nous nous focalisons ensuite sur le réseau d'eau qui fournit les données de consommation analysées dans cette thèse.

1.1.1 Réseau d'eau intelligent (smart grid)

Un réseau d'eau intelligent (*smart water grid*) (WERBOS, 2011) est un ensemble de solutions et de systèmes utilisant les nouvelles technologies de l'information et de la communication. Il permet aux opérateurs de réseaux d'eau de contrôler et diagnostiquer les problèmes. La figure 1.1 montre les différentes composantes d'un réseau d'eau intelligent, en partant des ressources en eau jusqu'à l'analyse et la mise à disposition des informations pertinentes pour répondre aux besoins des opérateurs et des consommateurs.

Dans un réseau d'eau intelligent, chaque bâtiment est équipé d'un ou de plusieurs compteurs communicants. Ces derniers transfèrent, par la voie de transmission radio, des informations concernant la consommation des usagers. Ces informations sont enregistrées dans les grandes bases de données qui sont accessibles aux opérateurs. Ces nouvelles solutions et systèmes sont ensuite couplés à des outils d'aide à la décision et de communication permettant d'optimiser :

- la gestion des réseaux (surveillance des réseaux à distance en temps réel, relève des compteurs à distance, etc.) ;
- la gestion de la ressource (accès à une information instantanée en cas de suspicion de fuite, adaptation des traitements de l'eau aux conditions météorologiques ou environnementales, etc.) ;
- le service aux consommateurs (facturation réelle trimestrielle, optimisation de la consommation des consommateurs, anticipation du budget d'eau, accès à l'historique des consommations en temps réel).

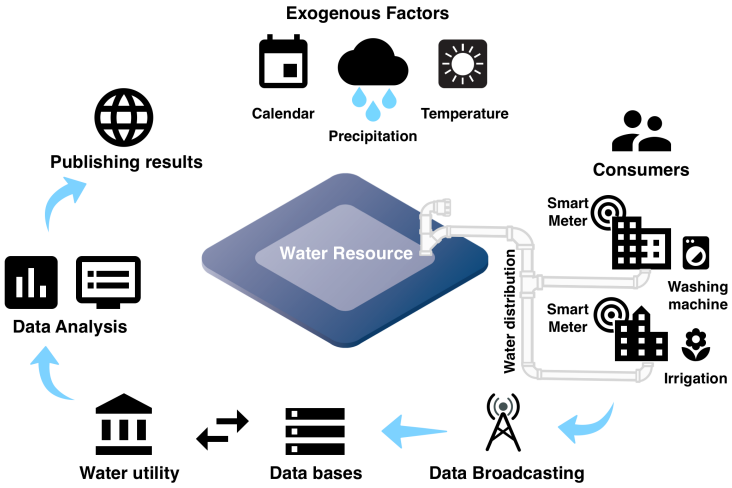


FIGURE 1.1 – Réseau d'eau intelligent et ses composantes

1.1.2 Compteur communicant

Un compteur communicant appelé *Smart Meter* (MAYER et collab., 1999) est un appareil électrique permettant de suivre à distance la consommation de manière détaillée et précise (voir figure 1.2). Pour ce faire, les compteurs sont dotés d'une technologie dite AMR (*automatic meter reading*) autorisant une lecture automatique de données et d'un module radio.

En utilisant cette technologie, les données de consommation peuvent être relevées à une fréquence élevée (de quelques minutes à quelques secondes). Le traitement de ces données peut conduire à des analyses très fines de la consommation, par exemple, la désagrégation de consommation proposée par COLE et STEWART (2013); FROEHLICH et collab. (2010) et la modélisation personnalisée du comportement des consommateurs proposée par FIELDING et collab. (2013).

FIGURE 1.2 – Compteur communicant installé au niveau de chaque bâtiment permettant de transférer les données de consommation

1.1.3 Acquisition des données

Le système de télérelevé de la consommation d'eau, appelé *Téléo*, est déployé sur le territoire du syndicat des eaux d'Île-de-France (SEDIF). Au 31 décembre 2015, on comptait près de 580 000 compteurs équipés d'un module radio sur l'ensemble des 149 communes du SEDIF (voir figure 1.3). Ils envoient, deux fois par jour pendant quelques secondes, une onde de très faible puissance, captée par des répéteurs disposés sur les toits d'immeubles, de candélabres et de mobiliers urbains. Cette onde et l'index du compteur sont transmis au centre de relation client de Veolia.

Grâce au télérelevé, la facture d'eau sera établie sur la consommation réelle et il est possible pour les usagers d'eau de consulter l'évolution de leur consommation dans leur espace client et de détecter ainsi d'éventuelles fuites sur leur réseau intérieur.

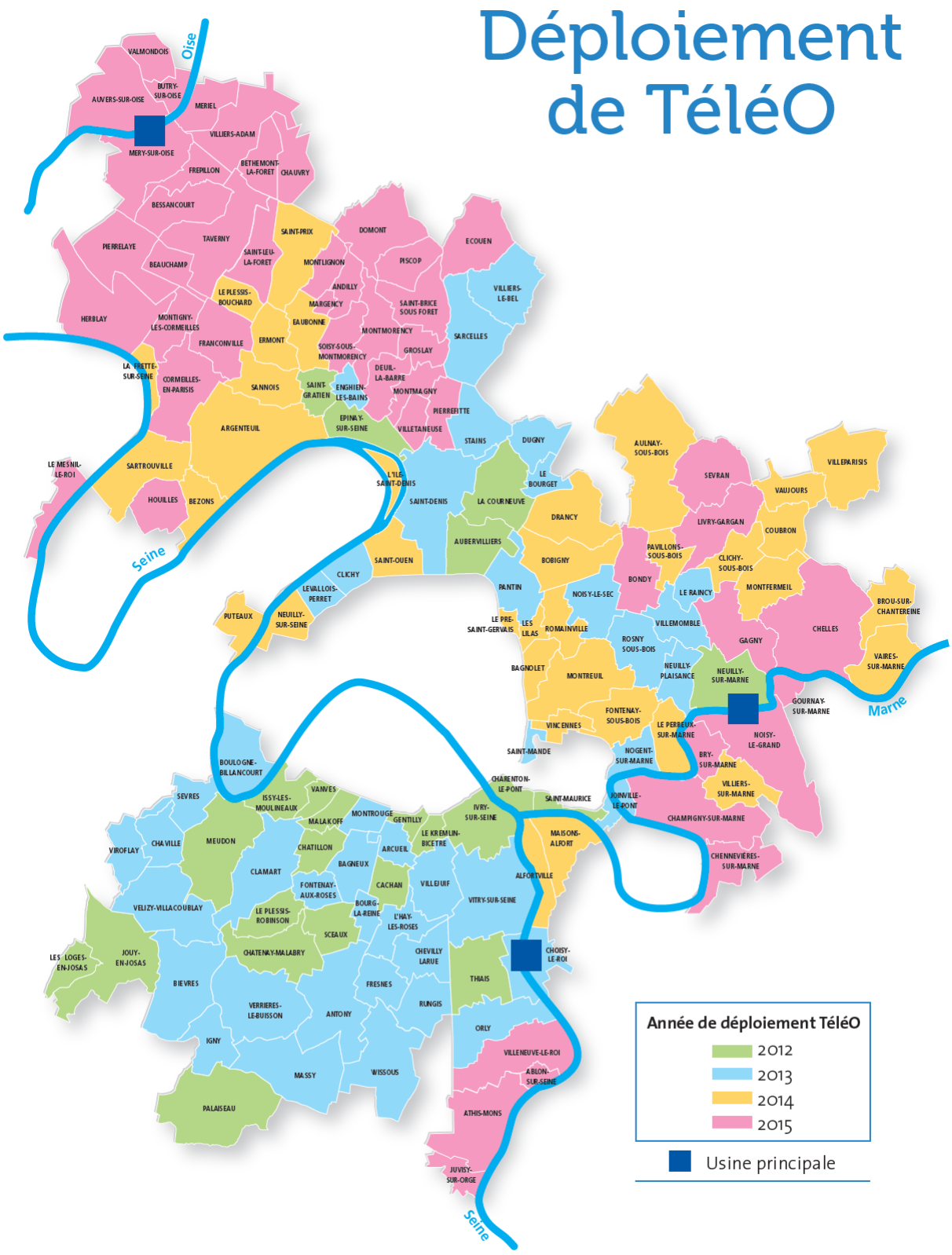


FIGURE 1.3 – Territoire du syndicat des eaux d'Ile-de-France (SEDIF)

Les données exploitées dans cette thèse sont les indexes de consommation croissants, à fréquence horaire, correspondant à un ensemble de 2 000 compteurs dites *Très Grands Consomma-*

teurs (TGC) répartis sur le territoire de SEDIF. Ceux-ci couvrent une période de 19 mois du 26 janvier 2015 au 24 juillet 2016. Les volumes de consommation horaires sont obtenus de la différence des indexes successifs.

1.2 État de l'art sur l'acquisition des données

La collecte des données de consommation est l'étape primordiale pour pouvoir mener les analyses et adopter des stratégies de gestion. On peut regrouper les données de consommation en deux groupes : faible résolution et haute résolution. Suivant la résolution des données, différents types d'analyse peuvent être menés.

Dans le cas de données de faible résolution, le recueil de données de consommation est généralement effectué tous les quatre mois (BRITTON et collab., 2008). Ces données peuvent être utilisées pour une planification à moyen ou à long terme de la consommation au niveau régionale. En particulier, elles sont utilisées pour étudier l'influence des variables socioéconomiques et de la saisonnalité sur la consommation (HOUSE-PETERS et CHANG, 2011; THOMAS et SYME, 1988). Ces approches sont basées, de manière générale, sur les modèles économétriques et les modèles de séries temporelles telle que la régression multivariée et exigent peu de données et de faibles ressources de calcul. Leur principal inconvénient est qu'elles ne permettent pas de représenter l'hétérogénéité temporelle de la consommation des habitats résidentiels qui requiert une résolution plus élevée des données collectées.

Avec l'avancée de la technologie, les compteurs communicants permettent aujourd'hui une collecte de données avec une résolution très élevée. Cela permet désormais de mieux caractériser la consommation au niveau résidentiel. Dans ce domaine, on peut également distinguer deux approches : la première consiste à installer des capteurs sur chaque appareil susceptible de consommer de l'eau (lave-linge, lave-vaisselle, douche, etc.) (ROWLANDS et collab., 2015) et la deuxième consiste à transmettre la consommation totale d'un résident au fil du temps. La première approche n'est pas généralement applicable dans les situations réelles car elle demande beaucoup de ressources et exige des coûts élevés. Au contraire, la deuxième approche représente une alternative plus acceptable (MAYER et collab., 1999).

Parmi les différents défis de recherche concernant la collecte de données, on peut citer :

- l'amélioration de la qualité du transfert à distance des données de consommation (au niveau logiciel et matériel dédié) (STEWART et collab., 2010) ;
- le développement d'un système de distribution de l'information qui peut enregistrer la masse de données et vérifier l'exactitude de celles-ci (ORACLE, 2009) ;
- l'analyse de l'impact de ce type d'équipement sur la vie privée des consommateurs (MCINTYRE, 2008).

1.3 Problématique et objectifs de la thèse

Bien que de nombreuses entreprises du secteur de l'énergie et de l'eau commencent à percevoir l'importance des données télérelevées des compteurs, celles-ci sont très souvent confrontées à des défis majeurs concernant la gestion et le traitement des données massives recueillies. D'un point de vue technique, les outils d'aide à la décision développés rencontrent des difficultés d'adaptation à la situation nouvelle. D'autre part, ces outils se focalisent sur le niveau de consommation prélevé par les compteurs, qui n'est pas toujours représentatif du comportement des usagers. Par exemple, dans un réseau d'eau, le niveau de consommation ne permet pas de différencier deux consommateurs qui présentent le même volume consommé mais qui ont deux habitudes de consommation différentes. En outre, l'efficacité des systèmes de gestion pourrait être impactée

par les changements dans le comportement de consommation des usagers d'un réseau. Ces changements, qui sont liés à des contextes différents (changement climatique, changement de saison, changement démographique) (KENNEY et collab., 2008), ne sont pas toujours faciles à détecter. Il est également difficile de caractériser la signature des changements et d'analyser leurs impact potentiel sur le réseau.

Cette thèse s'intéresse à la thématique *Data Analytics* des Smart Grids. L'objectif visé est de proposer des méthodologies génériques permettant d'analyser les masses de données recueillies via les compteurs intelligents et d'en extraire de l'information utile pour différents acteurs tels que les compagnies d'eau et d'électricité, les villes et les consommateurs.

Dans cette thèse, nous nous focalisons plus particulièrement sur l'analyse du comportement des consommateurs d'eau. Pour pouvoir mener à bien cette analyse, les méthodes développées doivent être en mesure de traiter la masse de données collectées. Pour atteindre cet objectif, une solution consiste à résumer les profils de consommation par des labels d'habitudes de consommation. Plusieurs méthodes permettent de faire cette discrétisation (KEOGH et collab., 2001; LIN et collab., 2003) ; ce qui permettrait également de suivre plus facilement le comportement des usagers dans le temps.

Dans une optique de développement d'un outil d'aide à la décision, il est souvent d'usage de regrouper les consommateurs en classes homogènes (FLATH et collab., 2012; HABEN et collab., 2015). Cette agrégation permettrait d'optimiser l'analyse du comportement à travers des groupes constitués. Ainsi, pour chaque groupe de consommateurs, suivant leur mode d'usage, une stratégie adaptée peut être mise en place. Par exemple, une stratégie d'incitation à de meilleures pratiques ou de tarification particulière peut être considérée pour un groupe présentant un niveau de consommation élevé.

La préservation de l'eau et la réduction des rejets d'eaux usées font partie également des objectifs majeurs des programmes environnementaux. Pour atteindre ces objectifs, des indications quant aux futures habitudes de consommation peuvent aider les opérateurs du secteur à mettre en place des programmes de préservation. Cela leur permettrait d'accroître la qualité des services proposés et de surmonter les problèmes liés à la pénurie d'eau potable. Dans ce cas, les méthodes de prévision du domaine d'apprentissage automatique peuvent être utilisées HARRINGTON (2012).

Le comportement des usagers évolue également au fil du temps. Cette évolution pourrait être de natures différentes (par exemple un changement ponctuel dû au passage des jours ouvrés au weekend). Dans cette thèse, on s'intéresse notamment à la détection de *changement structurel* (AUE et HORVÁTH, 2013) commun à un ensemble de compteurs. Travailler sur un ensemble de compteurs permet de détecter les changements de comportement qui ne seraient pas identifiables en considérant les compteurs individuels (CHEN et collab., 2013). L'objectif visé par les méthodes de détection de changement est de réduire le délai de détection et de pouvoir caractériser les changements détectés. Ainsi, en fonction du caractère des changements détectés, les mesures nécessaires peuvent être prises.

Les contributions de cette thèse visent donc à répondre aux problématiques suivantes :

- Réduire la taille des données issues d'un réseau d'eau potable pour pouvoir se focaliser sur les habitudes de consommation plutôt que sur le niveau d'eau consommé ;
- Proposer une méthode pour la classification automatique des consommateurs suivant la dynamique de leurs habitudes de consommation ;
- Prédire les futures habitudes de consommation au sein des groupes de consommateurs ;
- Détecter les changements structurels communs aux habitudes de consommation d'un ensemble d'usagers.

La spécificité des travaux réside dans l'approche générique adoptée pour le développement des outils d'aide à la décision, qui est basée sur les modèles dynamiques à variables latentes.

1.4 Organisation de la thèse

Cette section décrit l'organisation des chapitres du mémoire par ordre d'apparition.

Le chapitre 1 commence par une introduction générale concernant le contexte étudié dans cette thèse et décrit les problématiques associées, ainsi que les objectifs visés dans cette thèse. Les détails techniques concernant les réseaux d'eau intelligents et les compteurs d'eau sont introduits dans la suite. Ces derniers sont la base de l'acquisition des données de consommation analysées.

Le chapitre 2 présente dans un premier temps les données de consommation issues des compteurs intelligents. Ensuite, les prétraitements permettant de transformer ces données en habitudes journalières et hebdomadaires de consommation sont décrits. Les méthodes du domaine de la classification automatique de données fonctionnelles, notamment les approches basées sur le mélange de lois, sont exploitées. Les problèmes d'estimation des paramètres et de choix du modèle sont également passés en revue. Le passage de données continues à des données catégorielles permet de synthétiser les profils de consommation et de suivre plus facilement l'évolution du comportement des usagers du réseau d'eau. Finalement, deux bases de données catégorielles issues de l'étape de discrétisation résument le comportement de l'ensemble des compteurs.

Les bases de données catégorielles présentant l'évolution des habitudes de consommation de l'ensemble des compteurs sont exploitées dans le chapitre 3. Ce chapitre commence par un état de l'art sur les approches permettant de modéliser des séquences catégorielles. Ensuite, un algorithme de classification basé sur le mélange de modèles de Markov non homogènes est proposé pour regrouper les séquences catégorielles en classes; chaque classe étant caractérisée par son évolution markovienne propre au fil du temps. L'application de ce modèle sur les bases de compteurs permet d'identifier les groupes présentant une évolution similaire d'habitudes de consommation. Ce modèle permet de prédire les futures habitudes de consommation des consommateurs au sein des classes obtenues. Le chapitre s'achève par une comparaison, en termes d'erreurs de prévision, de la méthode proposée avec d'autres méthodes de l'état de l'art.

Le chapitre 4 est consacré au problème de la détection de changements communs à un ensemble de séquences catégorielles. Les données catégorielles et les groupes de compteurs obtenus dans les chapitres 2 et 3 sont exploités dans ce chapitre. Les méthodes séquentielles permettant de détecter des points de changement sont d'abord passées en revue. Ensuite, une méthode basée sur les tests séquentiels du rapport de vraisemblance est proposée pour la détection de changements communs à un ensemble de compteurs. Cette dernière est fondée sur les modèles de Markov non homogènes, qui permettent de modéliser la dynamique conjointe des habitudes de consommation avant et après les points de changement. La méthode proposée est évaluée sur les données synthétiques et puis appliquée sur les données réelles de consommation où les changements détectés sont interprétés.

Enfin, le manuscrit s'achève, dans le chapitre 5, par une conclusion et explore les pistes futures de recherche.

Chapitre 2

Données et prétraitements

Sommaire

2.1 Introduction

Dans ce chapitre, nous présentons d'abord les données de consommation d'eau potable collectées à une fréquence horaire à l'aide de compteurs d'eau communicants (voir chapitre 1).

Pour pouvoir traiter la masse de données recueillies, ce chapitre a pour objectif de synthétiser les profils journaliers et les profils hebdomadaires de l'ensemble des compteurs. Cela permet de réduire la quantité des données et de suivre plus facilement l'évolution des habitudes de consommation des consommateurs au fil du temps. Pour ce faire, une approche de classification automatique de données fonctionnelles, à savoir le mélange de régressions de Fourier (appelé en anglais *Fourier regression mixture model*) (FReMix) est utilisée pour identifier les principales habitudes de consommation. Ce modèle tient compte de la périodicité (souvent présente) des séries étudiées.

Finalement, en remplaçant les profils saisonniers par leurs labels d'habitudes de consommation, deux bases de données catégorielles sont conçues. Celles-ci résument l'évolution des habitudes de consommation de l'ensemble des compteurs.

2.2 Description des données

Dans cette thèse, nous analysons les données de consommation d'eau qui sont présentées sous forme de séries temporelles. Les données de consommation initialement collectées par les compteurs communicants sont les indexes de consommation croissants qui sont enregistrés à une fréquence horaire.

La différence des indexes successifs fournit le volume d'eau consommé. Ainsi, la base de données analysée est constituée du volume horaire d'eau potable pour un ensemble de $n = 2\,000$ compteurs (TGC) répartis autour de Paris (réseau du SEDIF) pour une période de 456 jours (du 26 janvier 2015 au 24 juillet 2016). Ces compteurs sont principalement associés à des habitations individuelles et à des immeubles collectifs avec un niveau de consommation élevé.

Dans le cadre statistique, nous formalisons les données de cette base par n séries temporelles $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, où chaque série $\mathbf{y}_i = (y_{i1}, \ldots, y_{i\mathrm{H}})$ correspond aux consommations horaires relevées sur le compteur indexé par i sur une durée H.

La figure 2.1 montre un extrait des séries de consommation associées à 20 compteurs durant 4 semaines (28 jours). Sur cette figure, nous avons également affiché la courbe moyenne (en orange) qui montre la tendance globale des séries de consommation. Les valeurs de consommations sont indiquées en mètres-cubes et on note la fréquence horaire des séquences observées dans le temps. On peut observer la périodicité des profils journaliers qui se répètent chaque semaine.

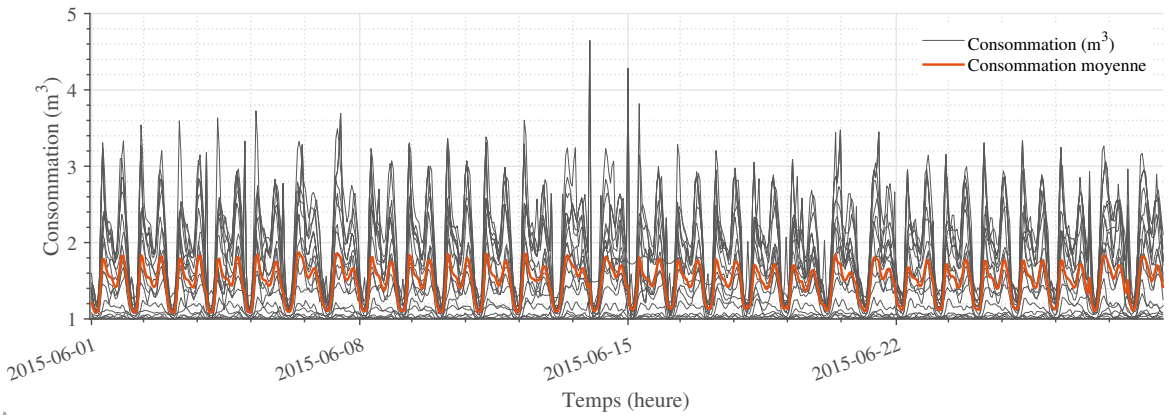
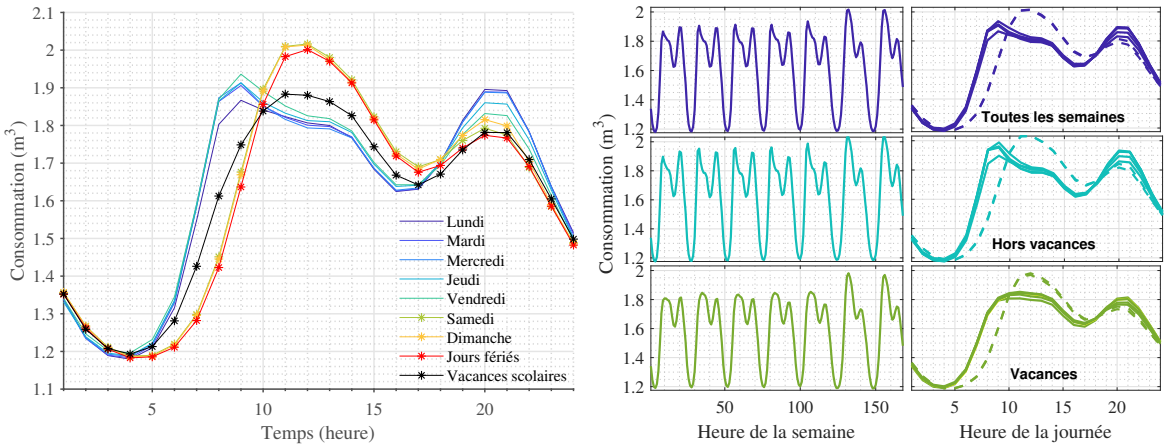


FIGURE 2.1 – Séries de consommation extraites de 20 compteurs durant une période de 4 semaines (du 1 juin 2015 au 29 juin 2015) et leur moyenne (en orange)



(a) Profils journaliers moyens

(b) Profils hebdomadaires moyens

FIGURE 2.2 – Courbes moyennes des profils journaliers (a) et des profils hebdomadaires (b)

Nous avons également calculé pour chaque type de jour (semaine, weekend, jours fériés, vacances scolaires), sa courbe de consommation qui est moyennée sur l'ensemble des compteurs (voir figure 2.2a). On peut observer que les courbes moyennes de consommation pour les jours ouvrés ont un comportement similaires entre elles. En revanche, elles présentent un comportement différent par rapport aux courbes de consommation de weekend. Le pic du matin est plus décalé et est plus prononcé pendant le weekend et leur pic du soir est moins élevé. Les jours fériés ont un comportement quasi similaire que celui du weekend. La courbe moyenne des vacances scolaires est caractérisé par un pic du matin diffus qui est moins prononcé que celui du weekend et plus décalé que celui de la semaine.

Dans la figure 2.2b, nous avons affiché les courbes moyennes hebdomadaires. Les courbes moyennes hebdomadaires sont affichées sur la partie gauche du graphique et à droite, nous avons présenté les profils journaliers de ces courbes, superposés. Dans ces graphiques, les pointillés désignent le comportement du weekend. On peut remarquer une similarité entre les courbes moyennes calculées sur toute la période et celles calculées en dehors de la période des vacances scolaires. En revanche, on note un comportement différent pendant les vacances scolaires, qui est caractérisé par un pic du matin plus diffus. Le comportement pendant le weekend reste quasiment identique.

2.3 Classification automatique de données

L'objectif visé ici est de regrouper les profils journaliers et hebdomadaires en classes, puis de résumer ces profils par leurs labels. Nous présentons d'abord une généralité des méthodes de classification automatique. Ensuite, la méthode fonctionnelle adoptée pour la discrétisation des courbes de consommation est introduite.

2.3.1 Généralités sur les méthodes de classification automatique

Les méthodes de classification automatique ont pour objectif de regrouper en classes des objets présentant des caractéristiques similaires. Celles-ci doivent être adaptées en fonction de la nature des données observées (données continues, catégorielles, binaires, fonctionnelles, etc.). Les méthodes utilisées dans ce domaine peuvent être regroupées en deux catégories : les méthodes non probabilistes et les méthodes probabilistes.

Les méthodes non probabilistes sont généralement basées sur une mesure de similarité pour le regroupement des objets. Elles ne reposent pas sur une mesure de probabilité ou d'incertitude

concernant le regroupement. La méthode de classification hiérarchique (WARD JR, 1963), la méthode des centre mobiles (MACQUEEN et collab., 1967), la méthode de classification ascendante hiérarchique (ARABIE et DE SOETE, 1996), et la méthode *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) (ESTER et collab., 1996) font partie de cette famille d'approches.

Les méthodes probabilistes modélisent les données à l'aide de distributions de probabilité. Le modèle de mélange (MCLACHLAN et BASFORD, 1988), qui fait partie de cette famille d'approches, suppose que les données proviennent d'un nombre fini de distributions (par exemple la distribution gaussienne). Ce modèle, qui constitue la base de l'approche utilisée pour le regroupement des données fonctionnelles, est décrit dans la section suivante.

Modèle de mélange

Le modèle de mélange considère que les données $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ sont distribuées indépendamment suivant un mélange de K distributions. Il est défini comme suit :

$$f(\mathbf{x}_i; \theta) = \sum_{k=1}^{\mathrm{K}} p_k f_k(\mathbf{x}_i; \theta_k), \tag{2.1}$$

où $f_k(\cdot)$ est la distribution de probabilité des observations au sein d'une classe k avec le vecteur des paramètres associé θ_k. Une probabilité a priori p_k est associée à chaque distribution, où $p_k > 0$ et $\sum_k p_k = 1$. Le vecteur $\theta = (\theta_1, \ldots, \theta_{\mathrm{K}}, p_1, \ldots, p_{\mathrm{K}})$ constitue l'ensemble des paramètres du modèle à estimer.

Dans le cas particulier où les données sont continues, il est d'usage de supposer que les composantes de mélange $f(\cdot; \theta_g)$ sont des densités gaussiennes $\mathcal{N}(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. À titre d'exemple, nous avons généré les données bidimensionnelles à partir de trois distributions gaussiennes. Les vecteurs de moyennes et les matrices de covariances sont fournis ci-dessous :

$$\boldsymbol{\mu} = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3.5 \\ 4.5 \end{pmatrix}, \begin{pmatrix} -1 \\ -0.5 \end{pmatrix} \right\}, \tag{2.2}$$

$$\boldsymbol{\Sigma} = \left\{ \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} \right\}. \tag{2.3}$$

La figure 2.3a montre un ensemble de 5 000 points générés suivant les valeurs de paramètres indiquées. L'application du mélange de gaussiennes sur ces données a abouti à l'estimation de trois composantes représentées dans la figure 2.3b.

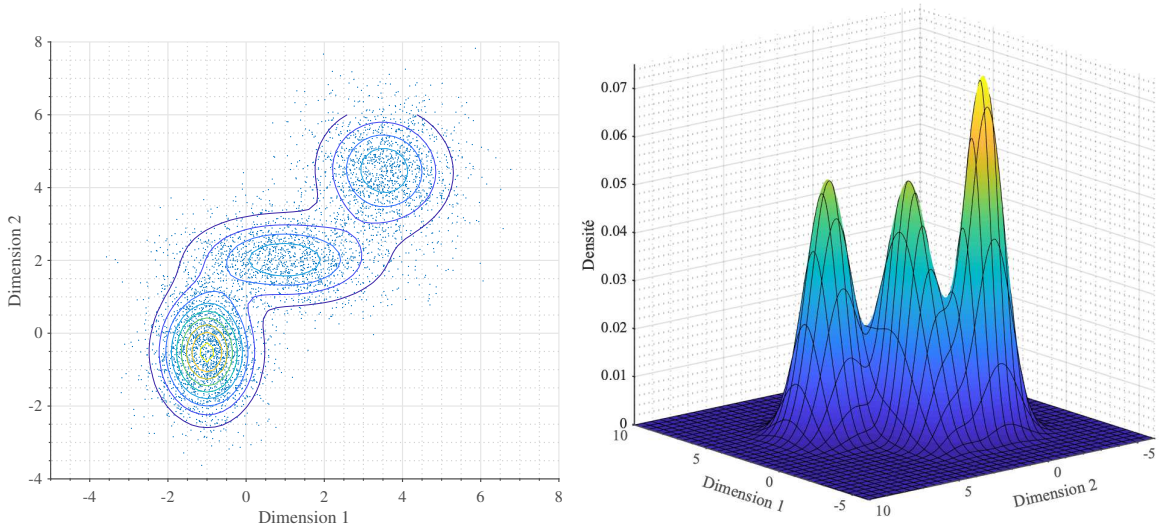
Le mélange de densités gaussiennes fait partie des approches les plus étudiées dans le domaine de la classification en raison de sa flexibilité et sa capacité à approcher une grande variété de densités (FRALEY et RAFTERY, 2002).

Estimation par maximum de vraisemblance et EM

Pour estimer les paramètres du modèle de mélange, la méthode généralement utilisée est celle du maximum de vraisemblance. Cette méthode permet d'estimer le paramètre θ en maximisant la fonction de vraisemblance $\mathcal{L}(\theta)$ suivante :

$$\mathcal{L}(\theta) = \log \mathrm{P}(\mathbf{x}; \theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{\mathrm{K}} p_k f(\mathbf{x}_i; \theta_k). \tag{2.4}$$

Cette fonction ne peut pas être maximisée de manière exacte mais des algorithmes itératifs permettant d'atteindre un maximum local peuvent être exploités. La méthode d'estimation la



(a) Données

(b) Densité (mélange de gaussiennes)

FIGURE 2.3 – Échantillon de 5000 points en deux dimensions généré à partir de trois gaussiennes bivariées (a) ; Densité de probabilité d'un mélange à trois composantes (b)

plus utilisée pour les modèles de mélange est l'algorithme Espérance-Maximisation (EM) (MCLACHLAN et KRISHNAN, 2007).

L'algorithme EM a été introduit pour la première fois par DEMPSTER et collab. (1977) pour estimer les paramètres d'un modèle comportant des données manquantes ou des variables latentes. Dans le cas des modèles de mélange, cette méthode tient compte de la variable latente intrinsèque au problème de classification (CELEUX et GOVAERT, 1992). On note $\mathbf{w}$, la variable latente qui gère le partitionnement des données. La log-vraisemblance des données complétées par les classes manquantes s'écrit de la manière suivante :

$$\mathcal{L}_c(\theta, \mathbf{w}) = \log \mathrm{P}(\mathbf{x}, \mathbf{w}; \theta_k) = \sum_{i=1}^{n} \sum_{k=1}^{\mathrm{K}} w_{ik} p_k f(\mathbf{x}_i; \theta_k) \tag{2.5}$$

où $w_{ik} = 1$ si $w_i = k$ et zéro sinon.

L'algorithme EM est une approche itérative qui commence avec une valeur initiale des paramètres $\theta^{(0)}$; à chaque itération de l'algorithme, les nouvelles valeurs des paramètres $\theta^{(q+1)}$ sont déduites à partir des paramètres précédents ($\theta^{(q)}$), de façon à maximiser l'espérance conditionnelle de la log-vraisemblance complétée qui est définie par

$$\begin{aligned} \mathrm{Q}(\theta; \theta^{(q)}) &= \mathbb{E}(\mathcal{L}_c(\theta, \mathbf{w}) | \mathbf{x}; \theta^{(q)}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{\mathrm{K}} \mathbb{E}(w_{ik} | \mathbf{x}_i; \theta^{(q)}) \log p_k f(\mathbf{x}_i; \theta_k^{(q)}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{\mathrm{K}} \mathbb{E}(w_i = k | \mathbf{x}_i; \theta^{(q)}) \log p_k f(\mathbf{x}_i; \theta_k^{(q)}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{\mathrm{K}} \tau_{ik}^{(q)} \log p_k f(\mathbf{x}_i; \theta_k^{(q)}), \end{aligned} \tag{2.6}$$

où

$$\tau_{ik}^{(q)} = p(w_i = k | \mathbf{x}_i; \theta^{(q)}) = \frac{p_k^{(q)} f(\mathbf{x}_i; \theta_k^{(q)})}{\sum_{h=1}^{\mathrm{K}} p_h^{(q)} f(\mathbf{x}_i; \theta_h^{(q)})} \tag{2.7}$$

désigne la probabilité a posteriori d'appartenance d'une observation $\mathbf{x}_i$ à une classe k, en se basant sur les paramètres $\theta^{(q)}$ de l'itération précédente.

L'algorithme EM répète les deux étapes suivantes jusqu'à la convergence vers un maximum local ou global de la log-vraisemblance :

- *Étape E* : indique l'étape de calcul de l'espérance Q et est réalisée par le calcul des probabilités a posteriori $\tau_{ik}^{(q)}$ d'appartenance d'une observation $\mathbf{x}_i$ à une classe k,
- *Étape M* : indique l'étape de la maximisation de la fonction Q qui est dédiée à la mise à jour des paramètres du modèle.

En pratique, plusieurs stratégies d'initialisation sont possibles (BIERNACKI et collab., 2003). L'une d'entre elles consiste à utiliser les *k-means*.

L'algorithme 1 décrit les deux étapes de l'algorithme EM pour estimer le vecteur des paramètres $\theta = (p_k, \boldsymbol{\mu}_k, \Sigma_k)_{1\leq k\leq K}$ d'un mélange de densités gaussiennes.

Algorithme 1 : Algorithme EM pour un mélange de lois gaussiennes

Entrées : n observations $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, nombre de composantes K, paramètre initial $\theta^{(0)} = (p_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \Sigma_k^{(0)})_{1\leq k\leq K}$

$q \leftarrow 0$;

répéter

Étape E : calcul des probabilités à posteriori :

$$\tau_{ik}^{(q)} = \frac{p_k^{(q)} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k^{(q)}, \Sigma_k^{(q)})}{\sum_{h=1}^{K} p_h^{(q)} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_h^{(q)}, \Sigma_h^{(q)})}$$

Étape M : mise à jour des paramètres :

$$\begin{aligned} p_k^{(q+1)} &\leftarrow (1/n) \sum_{i=1}^{n} \tau_{ik}^{(q)} \\ \boldsymbol{\mu}_k^{(q+1)} &\leftarrow \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{x}_i \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} \\ \Sigma_k^{(q+1)} &\leftarrow \sum_{i=1}^{n} \tau_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})' \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} \end{aligned}$$

$q \leftarrow q + 1$

jusqu'à ce que la vraisemblance converge;

Sorties : $\widehat{\theta}^{(q+1)} = (p_k^{(q+1)}, \boldsymbol{\mu}_k^{(q+1)}, \Sigma_k^{(q+1)})_{1\leq k\leq K}$

Une fois les paramètres estimés, on obtient une classification des données en attribuant chaque observation $\mathbf{x}_i$ à la classe k qui maximise la probabilité a posteriori τ_{ik}.

Algorithme CEM

L'algorithme Classification EM (CEM) initié par CELEUX et GOVAERT (1992), est un algorithme itératif qui estime simultanément les paramètres d'un modèle et le regroupement des observations. La principale différence avec l'algorithme EM réside dans le fait que cet algorithme maximise la vraisemblance complétée. L'algorithme CEM (de même que EM) commence avec une valeur initiale des paramètres $\theta^{(0)}$ et répète les trois étapes suivantes jusqu'à la convergence :

- *Étape E* : indique l'étape du calcul des probabilités a posteriori $\tau_{ik}^{(q)}$ (voir équation (2.7)) ;
- *Étape C* : indique l'étape d'affectation de chaque observation x_i à la classe $w_i^{(q)}$ qui maximise $\tau_{ik}^{(q)}, 1 \leq k \leq K$;

- *Étape M* : indique l'étape de la maximisation de la log-vraisemblance $\mathcal{L}_c(\theta, \mathbf{w}^{(q)})$ (voir équation (2.5)) qui est dédiée à la mise à jour des paramètres du modèle.

Cet algorithme inclut une étape de classification entre les étapes d'estimation de l'espérance et de maximisation issues de l'algorithme EM. Une extension en-ligne de cet algorithme est également proposée par par SAMÉ et collab. (2007).

Une généralisation des algorithmes EM et CEM est également proposée par AMBROISE et GOVAERT (2000) sous la forme d'une classification floue (appelé *fuzzy classification*) des observations. Pour trouver la meilleure partition (*fuzzy partition*), cet algorithme maximise un critère de vraisemblance pénalisé. Le terme de pénalisation est fondé sur l'entropie et mesure le degré d'incertitude (*fuzziness*) du regroupement des observations. Cet algorithme peut être considéré comme un compromis entre l'algorithme EM et CEM.

Sélection de modèle

Dans le cadre des modèles de mélange, plusieurs approches ont été proposées pour sélectionner le nombre de composantes. Parmi ces approches, on peut citer les critères basés sur l'entropie (CELEUX et SOROMENHO, 1996), les tests d'hypothèses (SOROMENHO, 1994), le facteur de Bayes (KASS et RAFTERY, 1995) et les critères d'information (CUTLER et WINDHAM, 1994).

Le principe de ces critères est de choisir le modèle qui fait croître le plus possible la vraisemblance, tout en minimisant la complexité du modèle. La plupart des critères se base sur le maximum de vraisemblance pénalisé par le nombre de paramètres libres du modèle ; ce qui conduit à l'expression générale suivante, à optimiser par rapport à différents modèles :

$$\mathrm{C}(\mathrm{K}) = \mathcal{L}(\hat{\theta}) - \lambda_c \vartheta(\mathrm{K}), \tag{2.8}$$

où $\hat{\theta}$ est l'estimation de maximum de vraisemblance pour le vecteur des paramètres θ et $\vartheta(\mathrm{K})$ désigne le nombre de paramètres du mélange. Le coefficient λ_c désigne le facteur de pénalisation de la complexité du modèle, qui est spécifique au critère C.

Le premier critère de sélection de modèle qui a été proposé par AKAIKE (1974) est connu sous le nom AIC (*Akaike Information Criterion*) :

$$\mathrm{AIC}(\mathrm{K}) = \mathcal{L}(\hat{\theta}) - 2\vartheta(\mathrm{K}). \tag{2.9}$$

BOZDOGAN (1987) a proposé une variante du critère d'Akaike, appelée AIC3, définie par :

$$\mathrm{AIC3}(\mathrm{K}) = \mathcal{L}(\hat{\theta}) - 3\vartheta(\mathrm{K}). \tag{2.10}$$

Le critère BIC (*Bayes information criterion*) (SCHWARZ et collab., 1978a) est utilisé dans les problèmes de sélection de modèle où la pénalité dépend également de la taille de l'échantillon :

$$\mathrm{BIC}(\mathrm{K}) = \mathcal{L}(\hat{\theta}) - \vartheta(\mathrm{K}) \log(n \times \mathrm{T}). \tag{2.11}$$

Le critère *Integrated Classification Likelihood* (ICL) proposé par (BIERNACKI et collab., 2000), est défini par

$$\mathrm{ICL}(\mathrm{K}) = \mathcal{L}_c(\hat{\theta}) - \vartheta(\mathrm{K}) \log(n \times \mathrm{T}). \tag{2.12}$$

Ces critères sont calculés pour différents nombres de classes K. Finalement, le modèle correspondant à la valeur optimale de critère est sélectionné.

2.3.2 Discrétisation des profils saisonniers de consommation et construction d'une base de données catégorielles

Dans le contexte des séries de consommation d'eau, nous nous focalisons ici sur l'analyse des habitudes de consommation (motifs) appelées également *états* plutôt que le niveau d'eau consommé; chaque habitude correspond ainsi à un comportement adopté par un consommateur au fil du temps. Les habitudes adoptées sont liées le plus souvent à un ensemble de facteurs. Parmi ces derniers, on peut citer le comportement inhérent, les évènements calendaires (CARDELL-OLIVER, 2013), les conditions météorologiques (GUTZLER et NIMS, 2005) ou bien un changement démographique (CAVANAGH et collab., 2002; HANKE et MARE, 1982; JONES et MORRIS, 1984). Dans l'optique de l'analyse temporelle des habitudes de consommation, nous proposons de discrétiser les profils saisonniers de consommation. L'objectif visé est d'une part de synthétiser les profils saisonniers de consommation pour faciliter l'interprétation des données massives recueillies, et d'autre part, d'analyser la dynamique de ces habitudes temporelles pour l'ensemble des compteurs.

On s'intéresse plus particulièrement à l'analyse de l'évolution des habitudes suivant deux visions : journalière et hebdomadaire (voir figure 2.4). Chacune de ces visions peut s'avérer nécessaire et peut aboutir à différents niveaux d'interprétations. La vision journalière permet de mener une analyse plus fine et plus locale des habitudes de consommation. Par exemple, cette vision pourrait conduire à une distinction entre deux habitudes de consommation en fonction de l'occurrence de leur pic du matin et du soir (voir figures 2.4c et 2.4d). En revanche, en utilisant cette vision locale, les informations qui sont observables à une échelle plus globale peuvent passer inaperçues. La vision hebdomadaire est ainsi importante, car elle permet de tenir compte des variations observées à l'échelle d'une semaine. Par exemple, elle permettrait de distinguer deux consommateurs en fonction de leur différence de comportement entre les jours ouvrés et le weekend (voir figures 2.4a et 2.4b).

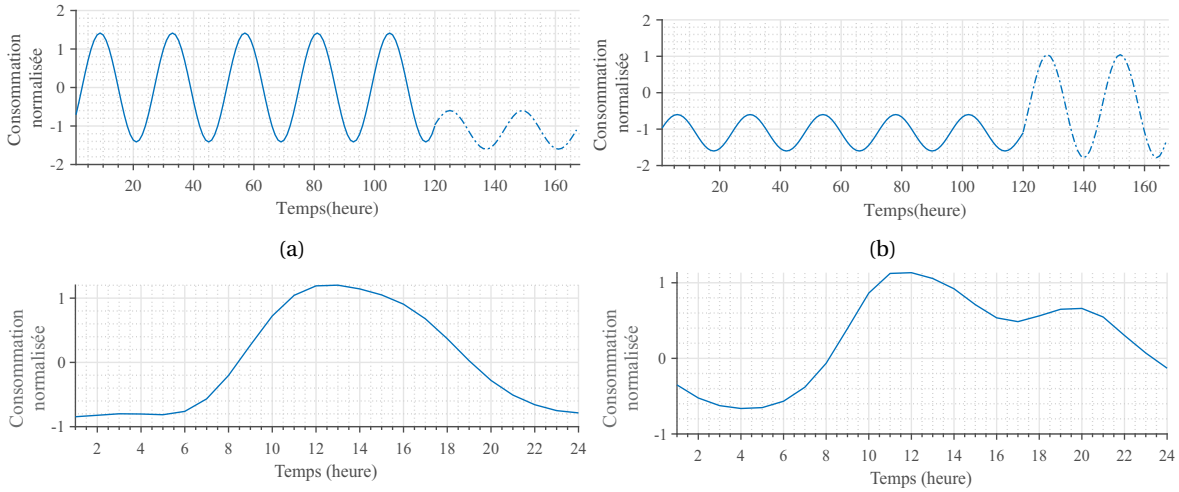


FIGURE 2.4 – Exemple illustratif (vision journalière et hebdomadaire) : deux profils hebdomadaires de consommation présentant un comportement opposé les jours ouvrés et le weekend (a) et (b); deux profils journaliers de consommation, l'un présentant un seul pic (c) et l'autre comportant deux pics (d).

Les étapes suivantes permettent de discrétiser les profils saisonniers de consommation (voir également le schéma 2.5) :

- découpage des séries en tranches saisonnières (journalières ou hebdomadaires) ; pour chaque série $\mathbf{y}_i$, on extrait S profils où S est le nombre de jours (S = 546) ou de semaines (S = 78),
- discrétisation des profils saisonniers à l'échelle du jour ou de la semaine par classification automatique dans le but de concevoir une base de données catégorielle représentant l'ensemble des habitudes de consommation de tous les compteurs,
- création d'une base de données catégorielles présentant la dynamique des habitudes de consommation de l'ensemble des compteurs.

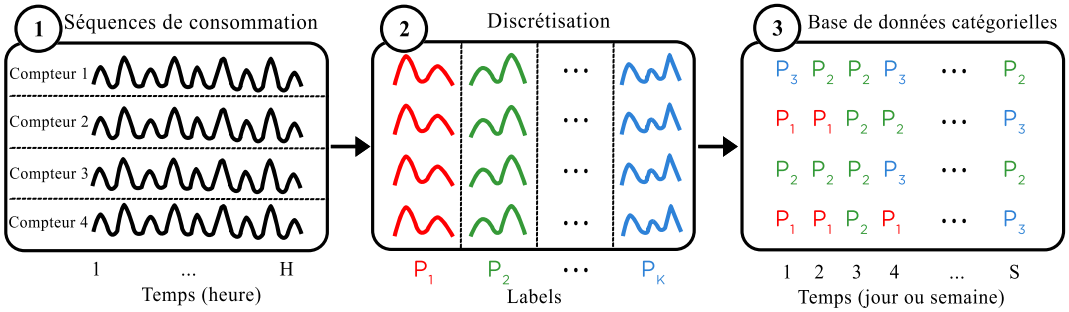


FIGURE 2.5 – Schéma indiquant la discrétisation des profils saisonniers : (1) : séries de consommation initiales observées à une fréquence horaire pendant H heures ; (2) : classification des profils saisonniers ; (3) : S habitudes (journalières ou hebdomadaires) de consommation associées.

Les 3 étapes mentionnées sont décrites dans les sections suivantes.

2.3.3 Extraction de profils saisonniers

Cette étape consiste à découper chaque série y_i en S tranches journalières ou hebdomadaires $(\mathbf{x}_{i1},\ldots,\mathbf{x}_{iS})$, avec $\mathbf{x}_{is} \in \mathbb{R}^{24}$ dans le cas des profils journaliers et $\mathbf{x}_{is} \in \mathbb{R}^{168}$ dans le cas des profils hebdomadaires. Afin de respecter l'hypothèse de gaussianité des modèles utilisés, les volumes de consommation sont préalablement remplacés par leur logarithme. La figure 2.6 montre un extrait des profils journaliers de consommation pour un ensemble de 20 compteurs et durant 12 jours. Dans cette figure, chaque colonne représente un jour de la semaine (lundi, mercredi, vendredi, dimanche). La figure 2.7 montre également les profils hebdomadaires de consommation associés aux mêmes compteurs portant sur 9 semaines. Dans cette figure, les dates sont celles du premier jour de la semaine (lundi) et la courbe de consommation moyenne est affichée en orange pour chaque graphique. L'ensemble de ces profils est utilisé comme entrées pour l'étape de classification.

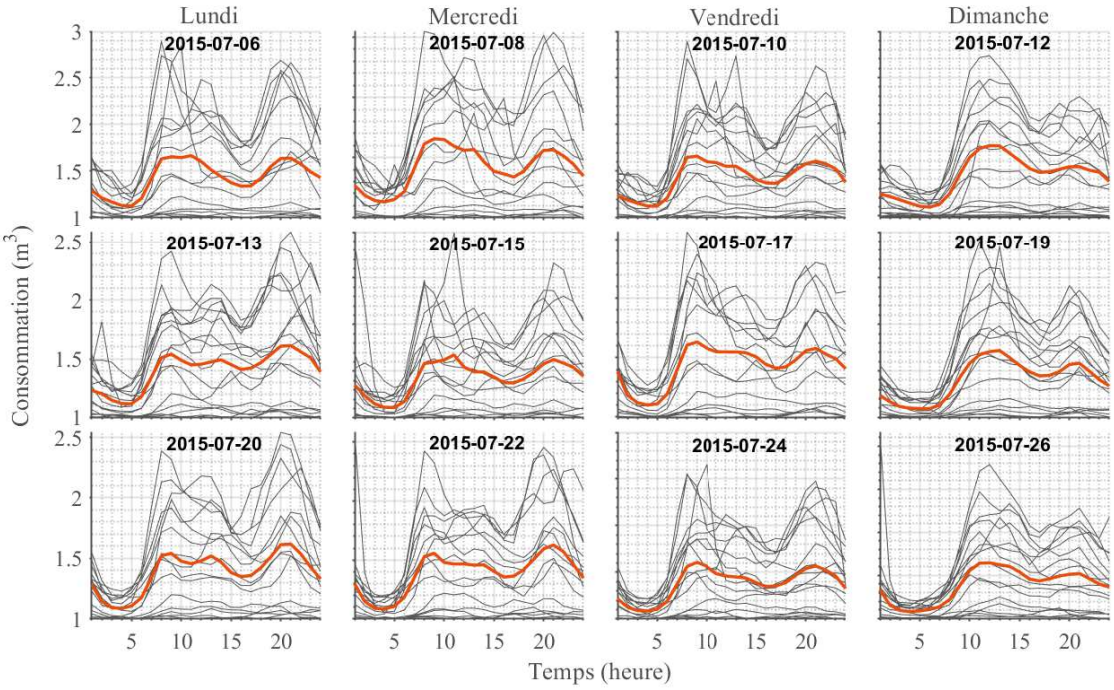


FIGURE 2.6 – Courbes de consommation journalières de 20 consommateurs associées aux 12 jours. Les graphiques dans chaque colonne sont liés, de gauche à droite, aux courbes de consommation de lundi, mercredi, vendredi et dimanche. La courbe de consommation moyenne est affichée en orange pour chaque graphique.

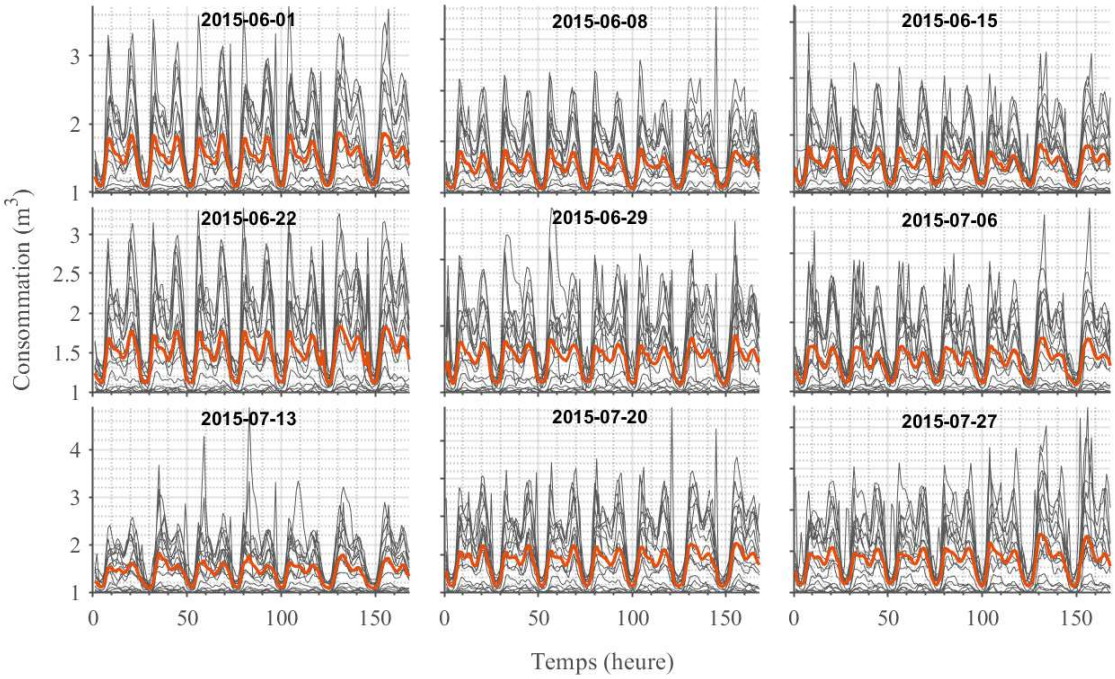


FIGURE 2.7 – Courbes de consommation hebdomadaires de 20 consommateurs associées aux 9 semaines. Les dates sur chaque graphique sont celles du premier jour de la semaine (lundi). La courbe de consommation moyenne est affichée en orange pour chaque graphique.

2.3.4 Classification de profils saisonniers

Cette étape consiste à résumer l'ensemble des courbes hebdomadaires ou journalières de consommation par un nombre réduit de profils. Compte tenu de la nature fonctionnelle de ces données, nous avons opté pour une méthode de classification qui tient compte de cette spécificité. Une extension du mélange de régressions polynomiales (GAFFNEY et SMYTH, 1999) proposée par SAMÉ et collab. (2016) et CHEIFETZ et collab. (2017) est utilisée. Dans ce dernier, les courbes polynomiales qui désignent les centres des classes, sont remplacées par des séries de Fourier. Ce modèle, appelé également FReMix (*Fourier regression mixture model*), a été proposé dans le cadre de classification des profils saisonniers issus d'un réseau d'eau. Ayant des classes à profils de consommation non linéaires ou périodiques, les polynômes de Fourier ont été préférés.

La densité de probabilité associée au modèle FReMix pour une série $\mathbf{x}_{is}$ est définie par

$$f(\mathbf{x}_{is};\theta) = \sum_{k=1}^{K} p_k \mathcal{N}(\mathbf{x}_{is};\mathbf{U}\boldsymbol{\zeta}_k,\sigma_k^2\mathbf{I}), \tag{2.13}$$

où $\theta = (p_1,\ldots,p_K,\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_K,\sigma_1^2\ldots,\sigma_K^2)$ est le vecteur des paramètres du modèle. Les p_k sont les proportions du mélange, qui vérifient $\sum_{k=1}^{K} p_k = 1$. Les $\boldsymbol{\zeta}_k \in \mathbb{R}^{2(q_1+q_2)}$ et σ_k^2 sont respectivement le vecteur des coefficients et la variance du bruit associés à la $k$ème composante du mélange. La matrice $\mathbf{U} = [\mathbf{u}_1,\ldots,\mathbf{u}_m]'$ est la matrice de régression de dimension $m \times 2(q_1+q_2)$, où $\forall\, t = 1,\ldots,m$, le vecteur $\mathbf{u}_t \in \mathbb{R}^{2(q_1+q_2)}$ est définie par

$$\begin{aligned}\mathbf{u}_t = \Big(&\cos(\frac{2\pi t}{24}) \;\; \sin(\frac{2\pi t}{24})\ldots\cos(\frac{2\pi q_1 t}{24}) \;\; \sin(\frac{2\pi q_1 t}{24}) \\ &\cos(\frac{2\pi t}{168}) \;\; \sin(\frac{2\pi t}{168})\ldots\cos(\frac{2\pi q_2 t}{168}) \;\; \sin(\frac{2\pi q_2 t}{168}) \Big)',\end{aligned} \tag{2.14}$$

et $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ est la densité gaussienne de moyenne $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$.

Le vecteur des paramètres du modèle est estimé en maximisant la log-vraisemblance

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{\mathrm{K}} p_k \mathcal{N}(\mathbf{x}_{is}; \mathbf{U}\boldsymbol{\zeta}_k, \sigma_k^2 \mathbf{I}) \tag{2.15}$$

via l'algorithme EM. L'algorithme proposé se distingue de l'algorithme EM gaussien classique par l'estimation des coefficients de régression $\boldsymbol{\zeta}_g$ associés aux séries de Fourier dans l'étape M. L'algorithme 2 détaille la procédure itérative d'estimation des paramètres du modèle.

Algorithme 2 : Algorithme EM pour le modèle FReMix

Entrées : $n \times \mathrm{S}$ séries $(\mathbf{x}_{1s}, \ldots, \mathbf{x}_{ns})_{1 \leq s \leq \mathrm{S}}$, nombre de composantes K, paramètre initial $\theta^{(0)}$

$q \leftarrow 0$;

répéter

Étape E : calcul des probabilités à posteriori :

$$\tau_{ik}^{(q)} = \frac{p_k^{(q)} \mathcal{N}(\mathbf{x}_{is}; \mathbf{U}\boldsymbol{\zeta}_k^{(q)}, \sigma_k^{2^{(q)}} \mathbf{I})}{\sum_{h=1}^{\mathrm{K}} p_h^{(q)} \mathcal{N}(\mathbf{x}_{is}; \mathbf{U}\boldsymbol{\zeta}_h^{(q)}, \sigma_h^{2^{(q)}} \mathbf{I})}$$

Étape M : mise à jour des paramètres :

$$\begin{aligned}
p_k^{(q+1)} &\leftarrow (1/n) \textstyle\sum_{i=1}^{n} \tau_{ik}^{(q)} \\
\boldsymbol{\zeta}_k^{(q+1)} &\leftarrow \left[\left(\textstyle\sum_{i=1}^{n} \tau_{ik}^{(q)} \right) \sum_{t=1}^{m} \mathbf{u}_t \mathbf{u}_t' \right]^{-1} \left[\textstyle\sum_{t=1}^{m} \left(\sum_{i=1}^{n} \tau_{ik}^{(q)} x_{it} \right) \mathbf{u}_t \right] \\
\sigma_k^{2^{(q+1)}} &\leftarrow \textstyle\sum_{i=1}^{n} \tau_{ik}^{(q)} \sum_{t=1}^{m} (x_{it} - \mathbf{u}_t' \boldsymbol{\zeta}_k^{(q+1)})^2 \Big/ \left(m \textstyle\sum_{i=1}^{n} \tau_{ik}^{(q)} \right)
\end{aligned}$$

$q \leftarrow q + 1$

jusqu'à ce que la vraisemblance converge;

Sorties : paramètre $\hat{\theta}$

Une fois que les paramètres du modèle ont été estimés, une partition des données est obtenue en affectant à chaque série $\mathbf{x}_i$ la classe g dont la probabilité a posteriori (voir étape E de l'algorithme) est maximale.

Pour améliorer le temps de calcul de cet algorithme, deux extensions présentées dans l'annexe C ont été utilisées :

- la première consiste à discrétiser les $n \times \mathrm{S}$ séries en 1 000 séries pondérées par l'algorithme de *k-means* (centres des classes obtenus avec K = 1 000), puis classifier ces nouvelles séries par une extension de l'algorithme FReMix qui prend en compte des pondérations dans les données. La réduction préalable des données via l'algorithme de *k-means* se comporte comme une méthode de quantification vectorielle (voir section C.1) ;
- la deuxième est basée sur le traitement distribué des données à l'aide des processeurs de type GPU (carte graphique) (voir section C.2).

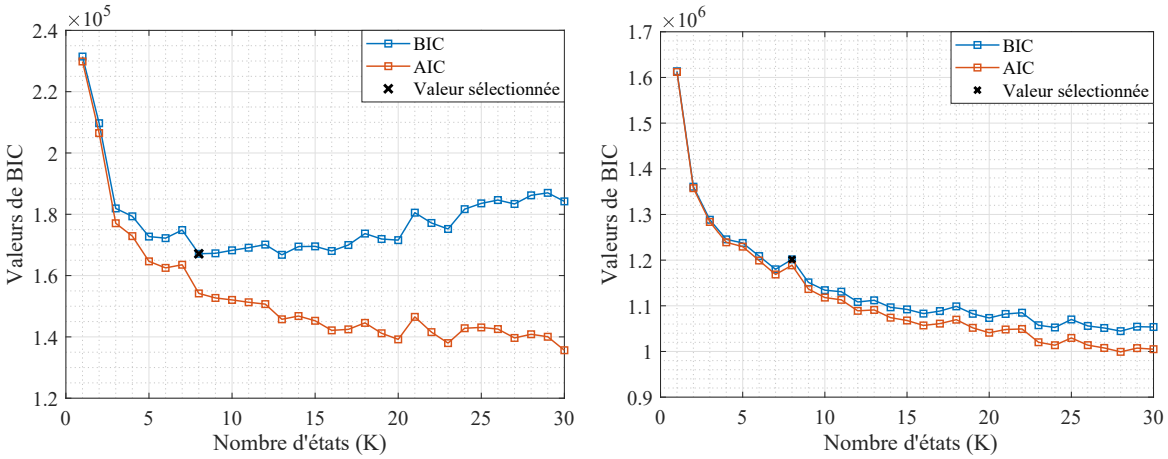
Choix des paramètres du modèle

Pour sélectionner le nombre optimal de classes ainsi que le nombre d'harmoniques q_1 et q_2 dans l'algorithme FReMix, Samé et collab. (2016) proposent une approche fondée sur le critère d'information de Bayes (BIC) défini par :

$$\mathrm{BIC}(\mathrm{K}) = -2\mathrm{L}(\hat{\boldsymbol{\theta}}) + \vartheta(\mathrm{K}) \log(n), \tag{2.16}$$

où $\hat{\boldsymbol{\theta}}$ est le vecteur des paramètres estimés par l'algorithme EM, et $\vartheta(K)$ désigne le nombre de paramètres libre pour un modèle constitué de K classes. Pour le modèle FReMix, nous avons $\vartheta(K) = 2K(q_1 + q_2 + 1) - 1$. Ce critère est calculé pour différentes valeurs de K, q_1 et q_2 et les valeurs de paramètres permettant d'obtenir la valeur minimale du critère est retenu.

Dans le cas de profils journaliers (voir figure 2.8a), on observe la remontée de la courbe du critère BIC à partir de K = 8, ce qui n'est pas le cas pour le critère AIC. Nous avons opté pour K = 8 états dans ce cas. Concernant les profils hebdomadaires de consommation (voir figure 2.8b), les deux critères montrent un comportement similaire avec des valeurs légèrement plus faibles pour le critère AIC. Nous observons également une première remontée des critères pour la valeur K = 8 et une décroissance après cette valeur. Dans ce cas, K = 8 nous a également semblé être un bon choix compte tenu de l'allure de BIC et suite à l'analyse des profils de consommation pour des nombres d'états inférieurs et supérieurs à 8. Concernant la paire (q_1, q_2), nous avons opté pour $(6,3)$ dans le cas de profils journaliers et $(4,24)$ dans le cas de profils hebdomadaires.



(a) Cas des profils journaliers

(b) Cas des profils hebdomadaires

FIGURE 2.8 – Critère BIC en fonction du nombre d'états K dans le cas de profils journaliers de consommation (a) et dans le cas de profils hebdomadaires de consommation (b)

Description des profils journaliers principaux (états)

La figure 2.9 montre les profils principaux obtenus en appliquant l'algorithme FReMix aux courbes journalières de consommation. Nous décrivons dans un premier temps la forme des profils-types (états) obtenus.

- État 1 et 7 : ces états sont caractérisés par un profil à 2 pics, où le pic du matin est plus élevé que celui du soir, et par un palier entre ces deux pics. Ces deux états peuvent représenter les habitudes de consommation de personnes actives ayant une activité professionnelle.
- État 2 et 6 : ces états représentent également un profil à deux pics où les pics sont plus étalés, et le pic du matin est décalé dans le temps.
- État 3 : cet état est représenté par un pic du matin très élevé par rapport à celui du soir qui est très diffus dans le temps.
- État 4 : cet état se caractérise par un profil à un seul pic qui est observé dans la journée. Cet état peut présenter l'habitude de consommation d'une zone d'activité.
- État 5 : est l'état avec la dispersion la plus élevée. Sa consommation moyenne est centrée autour de l'origine ; ce qui peut indiquer qu'il s'agit d'une classe de bruit.
- État 8 : cet état est représenté par un pic du matin très diffus et très faible par rapport à celui du soir. Celui-ci est également décalé dans le temps. Cet état peut présenter une habitude de consommation adopté pendant les vacances scolaires.

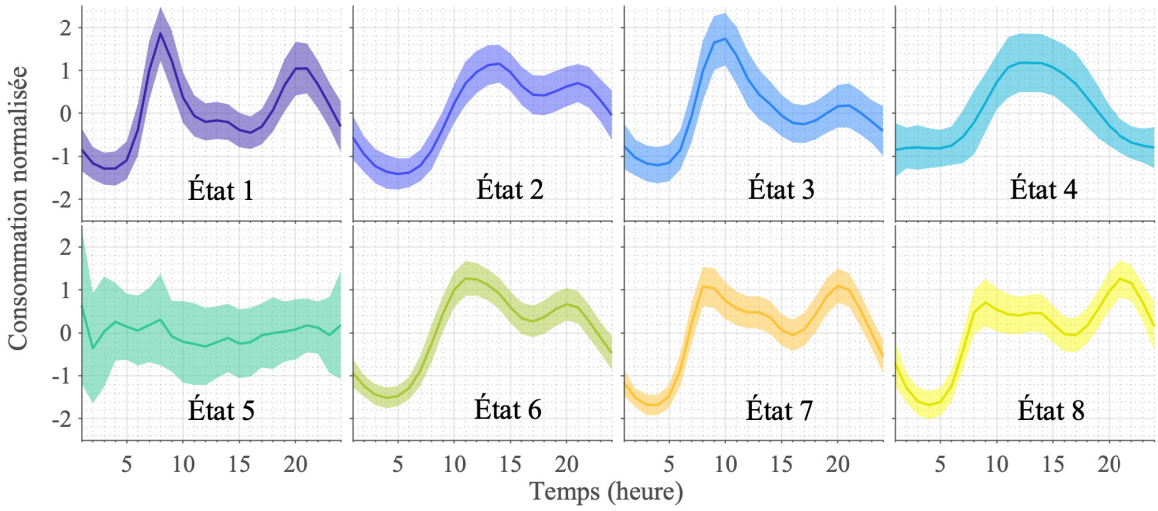
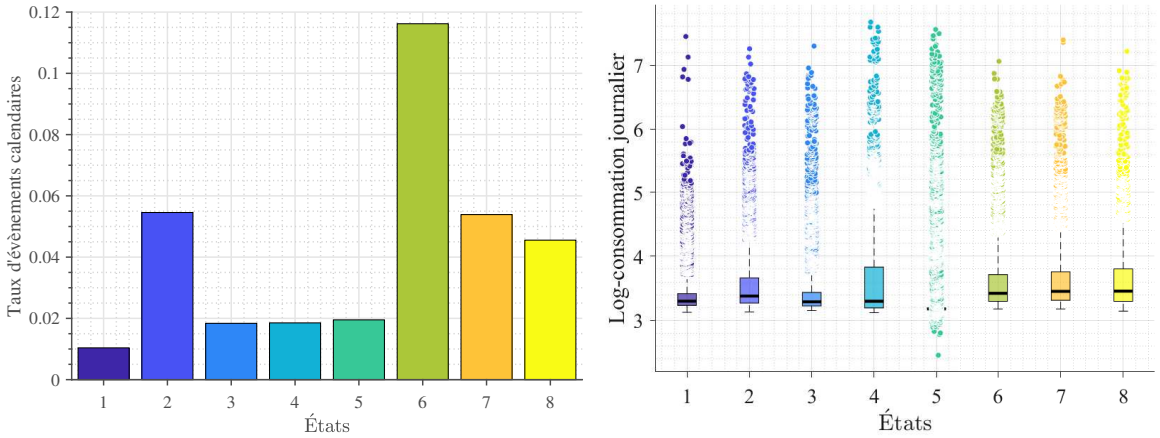


FIGURE 2.9 – 8 habitudes de consommations principales (appelé états) avec leur dispersion obtenues en regroupant les profils journaliers. Chaque état représente un centre d'une classe identifiée.

Nous avons également présenté le taux d'occurrence des évènements calendaires pour chaque état (voir figure 2.10a). On peut observer que les états 6, 2, 7 et 8 correspondent (par ordre décroissant de leurs occurrences) aux états les plus adoptés pendant les périodes de vacances scolaires. L'état 7 en est une exception, car cet état est fortement présent pendant toute la période.

Toujours dans l'optique de donner du sens aux profils-types d'usage, nous avons représenté la distribution des consommations réelles pour chaque état (voir figure 2.10b). On remarque une étendue plus élevée de la distribution de consommation pour les états 4 et 8, ce qui traduit le niveau d'eau consommé assez varié dans les zones d'activité et pendant les vacances d'été. Les états 1 et 3 ont en moyenne un volume de consommation relativement faible (médiane de consommation autour de 3,5 log-litres d'eau par jour). Concernant l'état 5 (bruit), on observe un nombre très élevé de valeurs extrêmes.



(a) Taux de vacances scolaires pour chaque état

(b) Distribution de la consommation par état

FIGURE 2.10 – Taux d'occurrence des vacances scolaires dans les états journaliers (a) et leur distribution du volume journalier de consommation (log-litres) (b)

Description des profils hebdomadaires principaux (états)

La figure 2.11 montre les profils principaux obtenus en appliquant l'algorithme FReMix aux profils hebdomadaires de consommation. Les profils hebdomadaires sont affichés avec leur dispersion sur la partie gauche du graphique et à droite, nous avons présenté les profils journaliers

superposés. Les pointillés désignent le comportement du weekend. Nous décrivons dans un premier temps la forme des profils-types (états) obtenus.

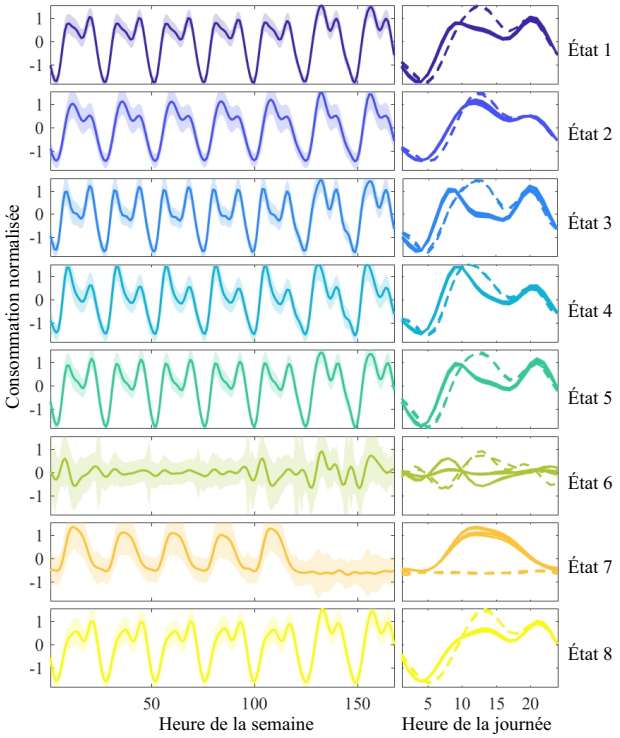
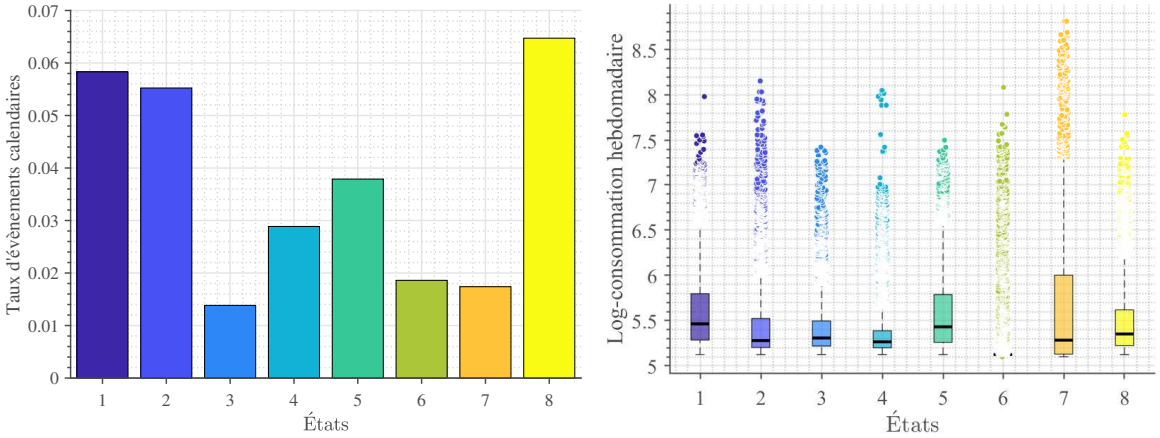


FIGURE 2.11 – 8 habitudes de consommation principales (appelé états) obtenues en regroupant les profils hebdomadaires. Chaque état représente un centre d'une classe identifiée : à gauche les profils hebdomadaire avec leur dispersion et à droite les profils journaliers superposés. Dans ces derniers, le weekend est représenté par les pointillés.

- État 1 : cet état a un comportement les jours ouvrés différent du weekend. Pendant les jours ouvrés, on observe un pic de consommation le matin qui est plus faible et qui est plus étalé dans le temps que celui du soir. On note également une décroissance lente du pic du matin. Le weekend, on observe deux pics, le premier étant plus élevé que le second ; le pic du matin est plus décalé dans le temps que celui des jours ouvrés. Cet état représente les habitudes de consommation de personnes actives ayant une activité professionnelle.
- État 2 : cet état est caractérisé par un comportement les jours ouvrés assez proche du comportement le weekend, ainsi que par deux pics (matin et soir) ; le pic du matin étant bien plus élevé que celui du soir. On n'observe pas de palier entre les deux pics mais plutôt une décroissance lente.
- État 3 et 5 : ces états présentent presque la même dynamique que l'état 1 ; la principale différence est l'apparition d'un palier entre les deux pics qui est plus marqué pour l'état 3.
- État 4 : cet état se caractérise par ses pics du matin plus élevés que ceux du soir et par les deux pics du weekend rapprochés.
- État 6 : le comportement de cet état pendant le weekend est en parfaite opposition (voir symétrie observée sur ses profils journaliers) avec son comportement le lundi et le vendredi. Le pic de consommation est atteint le weekend en début d'après midi et, de manière symétrique, la valeur minimale est atteinte aux mêmes heures le lundi et le vendredi. Du mardi au jeudi, la consommation reste relativement constante.
- État 7 : dans cet état, l'usage de l'eau potable durant les jours ouvrés est différent de son usage le weekend. En semaine, on observe une consommation quasi continue entre 10h et 20h et une consommation plus faible et constante le weekend. Cet état peut être associé à des bureaux, des administrations et des écoles.

- État 8 : cet état présente presque la même dynamique de consommation que les états 1, 3 et 5 pendant le weekend. La principale différence avec ces états se situe au niveau du pic de matin des jours ouvrés ; ce dernier est plus diffus et plus décalé dans le temps pour cet état.

Le taux d'occurrence des périodes de vacances scolaires pour les états hebdomadaires est affiché dans la figure 2.12a. Selon ce dernier, les états 8, 1, 2 et 5 correspondent (par ordre décroissant de leurs occurrences) aux états les plus présents pendant les périodes de vacances scolaires. L'état 1 en est une exception, car cet état est présent sur toute la période de consommation. L'état 3 possède le taux de vacances scolaires le plus faible. Cet état, de même que l'état 1 et 5, est associé à la catégorie « activité professionnelle ».



(a) Taux des vacances scolaires

(b) Distribution de la consommation

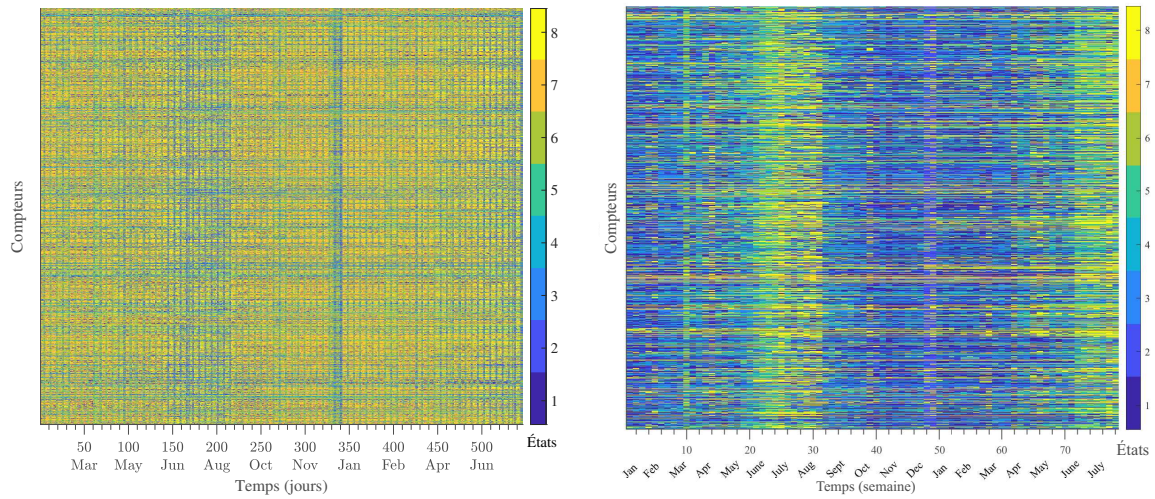
FIGURE 2.12 – Taux d'occurrence des vacances scolaires dans les états hebdomadaires (a) et leur distribution du volume hebdomadaire de consommation (log-litres) (b)

La figure 2.12b montre la distribution des consommations réelles pour chaque état hebdomadaire identifié. On constate une distribution étendue de la consommation pour l'état 7, ce qui induit une certaine variabilité de la consommation hebdomadaire pour cet état (représentant « administration »). Concernant les états 1, 5 et 8, on note un volume de consommation relativement élevé (médiane plus élevée). L'état 1 est présent sur toute la période et l'état 5 et 8 sont fortement présents pendant les périodes de vacances scolaires. Ce qui traduit une consommation plus élevée pendant les périodes de vacances scolaires.

2.3.5 Construction des bases de données catégorielles d'habitudes de consommation

Une fois que les profils-types (états) de consommation sont identifiés en utilisant l'ensemble des compteurs, on s'intéresse à leurs évolutions temporelles. En remplaçant les courbes journalières et hebdomadaires de consommation de chaque compteur par leurs labels d'habitudes de consommation obtenus via l'algorithme FReMix, deux bases de données sont constituées. Ces bases de données catégorielles sont représentées dans la figure 2.13. Dans cette figure, les lignes représentent les compteurs et les colonnes représentent l'évolution des états dans le temps (jour et semaine). Les états identifiés dans les figures 2.9 et 2.11 sont illustrés en utilisant le même code de couleurs pour chaque séquence.

Ces deux bases de données issues des états journaliers et hebdomadaires de consommation seront utilisées comme entrée pour les méthodes présentées dans le chapitre 3. L'objectif sera de modéliser la dynamique des habitudes de consommation des compteurs présents dans ces bases. On s'intéressera plus concrètement au regroupement des séquences d'habitudes de consommation et à la prévision de leurs futures habitudes de consommation.



(a) États journaliers

(b) États hebdomadaires

FIGURE 2.13 – Bases de données catégorielles obtenues à l'issue de l'étape de discrétisation de profils journaliers (a) et de profils hebdomadaires (b)

2.4 Conclusion

Ce chapitre a abordé le problème de classification des données fonctionnelles. Ces dernières font généralement l'objet d'une périodicité au fil du temps comme cela a été le cas des profils saisonniers de consommation d'eau analysés. Pour résoudre ce problème, en tenant compte de la périodicité des séries étudiées, nous avons opté pour le modèle de mélange de régressions Fourier (FReMix). Cette méthode permet également de traiter les séquences présentant des horizons temporels différents (jour, semaine, mois, etc.).

Dans le but de réduire la taille des données et de synthétiser les données de consommation, l'application de la méthode FReMix sur les profils saisonniers (journaliers et hebdomadaires) de consommation a conduit à l'identification de 8 habitudes de consommation principales ; chaque profil journalier ou hebdomadaire est remplacé ainsi par son label d'habitude de consommation correspondant. Les résultats obtenus ont également permis de suivre et d'interpréter plus facilement l'évolution des habitudes de consommation dans le temps.

Finalement, deux bases de données catégorielles incluant les habitudes journalières et hebdomadaires de consommation ont été conçues. Celles-ci résument l'évolution du comportement de l'ensemble des compteurs et feront l'objet d'une modélisation temporelle dans le chapitre suivant.

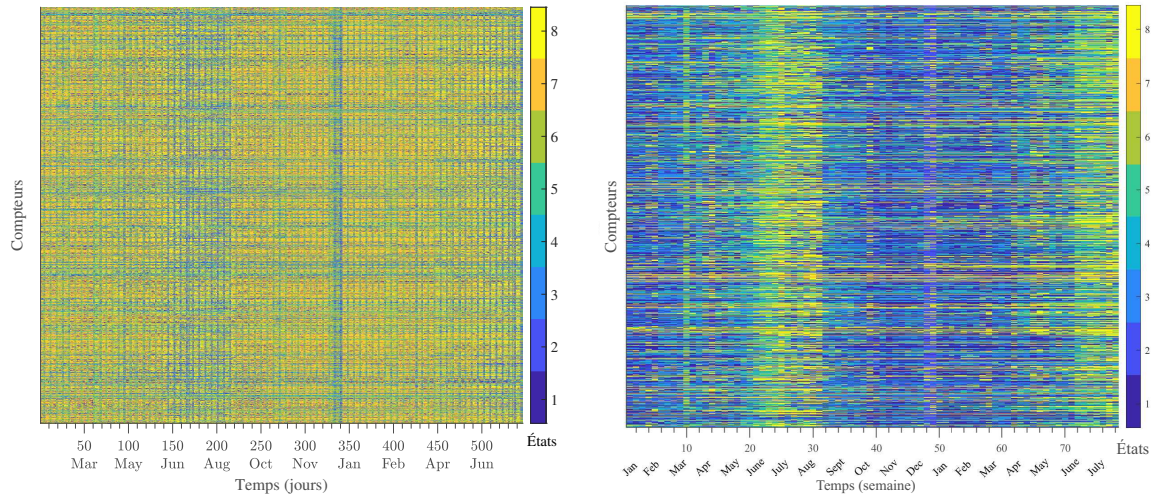
Chapitre 3

Classification et prévision de séquences catégorielles

Sommaire

3.1 Introduction

Dans le chapitre précédent, une étape de discrétisation des courbes de consommation a permis de concevoir deux bases de données catégorielles incorporant l'évolution des habitudes (journalières et hebdomadaires) de consommation de l'ensemble des compteurs (voir figure 3.1). Ce chapitre se focalise dans un premier temps sur le regroupement de ces séquences d'habitudes de consommation ; chaque groupe étant caractérisé par une évolution similaire de ses séquences dans le temps. Ensuite, l'objectif visé est de prédire les habitudes de consommation au sein des classes obtenues.



(a) États journaliers

(b) États hebdomadaires

FIGURE 3.1 – Bases de données catégorielles obtenues à l'issue de l'étape de discrétisation de profils journaliers (a) et de profils hebdomadaires (b)

Dans le cas des données de consommation d'eau potable, plusieurs études ont montré que certains facteurs exogènes peuvent impacter la dynamique des habitudes de consommation au fil du temps. Parmi ceux-ci, on peut citer les attitudes liées à l'environnement et à la préservation de l'eau (WILLIS et collab., 2011), aux variables climatiques (HOUSE-PETERS et collab., 2010; ZHOU et collab., 2000), aux informations socioéconomiques et démographiques (DOMENE et SAURÍ, 2006) ainsi qu'aux évènements calendaires (ZHOU et collab., 2002). La prise en compte de ces facteurs peut contribuer à une modélisation plus réaliste de l'évolution des habitudes de consommation.

Pour modéliser la dynamique conjointe de l'évolution des habitudes de consommation, nous proposons une méthodologie fondée sur un modèle de mélange ; chaque composante étant un modèle de Markov non homogène. La spécificité des modèles de Markov non homogènes se traduit par leur capacité à modéliser conjointement le comportement dynamique au fil du temps. Ce modèle permet dans un premier temps de regrouper les compteurs en classes ; chaque classe étant caractérisée par sa propre dynamique markovienne. Les facteurs de contexte peuvent également être utilisés comme variables d'entrée. Dans un second temps, en exploitant les paramètres estimés du modèle, les futures habitudes de consommation peuvent également être prédites au sein de chaque classe.

Dans la suite de ce chapitre, nous présentons d'abord un état de l'art portant sur les méthodes de classification et de prévision les plus utilisées. Ensuite, la méthodologie proposée est décrite et les résultats obtenus sont présentés.

3.1.1 Méthodes de classification de séquences catégorielles

Cette section commence par un état de l'art sur les méthodes de classification permettant de regrouper les séquences temporelles de nature catégorielle. Nous nous intéressons plus particulièrement aux méthodes basées sur des modèles probabilistes. Ensuite, quelques méthodologies utilisées dans le domaine des Smart Grids sont décrites.

Rappels sur les modèles de Markov

Les chaînes de Markov constituent un cadre adapté pour modéliser les séries temporelles. Une chaîne de Markov est un processus de Markov à temps discret, ou à temps continu et à espace d'états discret. Un processus de Markov à temps discret est une séquence de variables aléatoires à valeurs dans un espace d'états fini noté **E**. Une chaîne de Markov d'ordre 1 (voir figure 3.2) est définie par :

$$\mathrm{P}(z_t|z_{t-1}, z_{t-2}, \ldots, z_1) = \mathrm{P}(z_t|z_{t-1}) \ ; \ \forall t \geq 1, \tag{3.1}$$

où z_t désigne un état du modèle de Markov associé à l'instant t. Les chaînes de Markov d'ordre 1 supposent que chaque état à l'instant t ne dépend que de l'état à l'instant $t-1$. Dans la suite, on décrit respectivement les modèles de Markov homogènes et les modèles de Markov non homogènes.

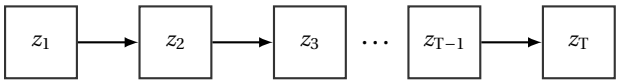


FIGURE 3.2 – Chaînes de Markov à temps discret

Chaînes de Markov homogènes

Les chaînes de Markov homogènes supposent que les probabilités de transition sont indépendantes de t :

$$\mathrm{P}(z_{t+1} = k|z_t = \ell) = \mathrm{P}(z_t = k|z_{t-1} = \ell), \ \ \forall t \geq 1, \ \forall (k,\ell) \in \mathbf{E}^2. \tag{3.2}$$

Une chaîne de Markov homogène est définie complètement par les probabilités initiales π_k et celles de transition $\mathbf{a}_{k\ell}$. Les probabilités de transition constituent une matrice de transition $(\mathbf{a}_{k,\ell})_{1\leq k,\ell\leq \mathrm{K}}$ de dimension $\mathrm{K} \times \mathrm{K}$ où K désigne le nombre d'états. Chaque élément de cette matrice peut être calculé en utilisant la formule suivante :

$$\mathbf{a}_{k,\ell} = \frac{\sum_{t=2}^{\mathrm{T}} z_{t,k}\, z_{t-1,\ell}}{\sum_{t=2}^{\mathrm{T}} \sum_{\ell'} z_{t,k}\, z_{t-1,\ell'}}, \tag{3.3}$$

où $\mathbf{a}_{k,\ell}$ désigne la probabilité de transition de l'état ℓ vers l'état k.

Chaînes de Markov non homogènes

Les chaînes de Markov non homogènes supposent que l'évolution des états est liée à des facteurs observés $\mathbf{u}_t$ (appelé également vecteur des descripteurs) à chaque pas de temps t :

$$\mathrm{P}(\mathbf{z}_t \,|\, \mathbf{z}_1, \ldots, \mathbf{z}_{t-1}, \mathbf{u}_1, \ldots, \mathbf{u}_{t-1}) = \mathrm{P}(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{u}_t). \tag{3.4}$$

Ainsi, une matrice de transition est calculée pour chaque valeur de t. Ce modèle est défini par :

$$f(\mathbf{z}|\mathbf{u}) = \mathrm{P}(z_1|u_1)\prod_{t=2}^{\mathrm{T}}\mathrm{P}(z_t|z_{t-1},\mathbf{u}_t). \tag{3.5}$$

Les probabilités initiales et celles de transition sont généralement formalisées en utilisant un modèle de régression logistique multinomiale :

$$\mathrm{P}(\mathbf{z}_1=\ell\mid\mathbf{u}_1) \quad = \quad \pi_\ell(\mathbf{u}_1;\boldsymbol{\alpha}) = \frac{e^{\boldsymbol{\alpha}_\ell^\top\mathbf{u}_1}}{\sum_{\ell'=1}^{\mathrm{K}}e^{\boldsymbol{\alpha}_{\ell'}^\top\mathbf{u}_1}} \tag{3.6}$$

$$\mathrm{P}(\mathbf{z}_t=k\mid\mathbf{z}_{t-1}=\ell,\mathbf{u}_t) \quad = \quad \pi_{k,\ell}(\mathbf{u}_t;\boldsymbol{\beta}_\ell) = \frac{e^{\boldsymbol{\beta}_{k,\ell}^\top\mathbf{u}_t}}{\sum_{\ell'=1}^{\mathrm{K}}e^{\boldsymbol{\beta}_{k,\ell'}^\top\mathbf{u}_t}}, \tag{3.7}$$

où π_ℓ désigne les probabilités initiales avec les paramètres $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_\ell)_{\ell=1,\ldots,\mathrm{K}}$ et $\pi_{k,\ell}$ désigne les probabilités de transition avec les paramètres $\boldsymbol{\beta}_\ell = (\boldsymbol{\beta}_{k,\ell})_{k=1,\ldots,\mathrm{K}}$. Ces probabilités dépendent d'un ensemble de variables exogènes désigné par $\mathbf{u}_t$.

La méthode du maximum de vraisemblance permet d'estimer l'ensemble des paramètres du modèle en maximisant :

$$\mathcal{L}(\theta) \quad = \quad \log\mathrm{P}(\mathbf{z}_t,\mathbf{u}_t;\theta) = \log\mathrm{P}(\mathbf{z}_1\mid\mathbf{u}_1;\theta) + \sum_{t=2}^{\mathrm{T}}\log\mathrm{P}(\mathbf{z}_t\mid\mathbf{z}_{t-1},\mathbf{u}_t;\theta). \tag{3.8}$$

Partitionnement basé sur les mélanges de modèles de Markov

Un mélange de modèles Markov proposé par Smyth (1999) permet de regrouper les observations selon leur caractéristique statique et leur caractéristique dynamique. On note $\mathbf{u}_i \in \mathbb{R}^m$ étant une variable aléatoire multivariée de dimension m et $\mathbf{z}_i$ étant une variable aléatoire en forme de séquence qui présentent respectivement les caractéristiques statiques et dynamiques d'une observation i. On note également $\mathbf{w} = (w_1,\ldots,w_n)$ étant une variable aléatoire associée à un ensemble de n observations qui gère leur partitionnement. Ce modèle est défini par :

$$f(\mathbf{z}_i,\mathbf{u}_i|\boldsymbol{\phi}) = \sum_{g=1}^{\mathrm{G}} p_g f_g(\mathbf{z}_i,\mathbf{u}_i|w_i=g;\boldsymbol{\phi}_g), \tag{3.9}$$

où G est le nombre de composantes du mélange, p_g désigne la probabilité marginale pour une classe g, $f(\mathbf{z},\mathbf{u}|\boldsymbol{\phi}_g)$ est la distribution conjointe des données observées appartenant à une classe g et $\boldsymbol{\phi}$ est l'ensemble des paramètres du modèle qui sera estimé de manière itérative à l'aide de l'algorithme EM. L'étape E de cet algorithme consiste à calculer la probabilité a posteriori d'appartenance d'une observation i à une classe g comme suit :

$$\mathrm{P}(w_i=g|\mathbf{z}_i,\mathbf{u}_i) = \frac{p_g f_g(\mathbf{z}_i,\mathbf{u}_i|w_i=g;\boldsymbol{\phi}_g)}{\sum_{g'=1}^{\mathrm{G}} p_{g'} f_{g'}(\mathbf{z}_i,\mathbf{u}_i|w_i=g';\boldsymbol{\phi}_{g'})}. \tag{3.10}$$

Une fois les probabilités a posteriori calculées, l'étape M de l'algorithme EM consiste à mettre à jour l'ensemble des paramètres du modèle. Cette méthode s'est révélée particulièrement pertinente pour le regroupement des utilisateurs de website selon leur comportement dynamique (accès aux pages Web) et en fonction de leur caractéristiques statiques (age et genre).

Classification en utilisant les modèles de classes latente

Le modèle des classes latentes (LCA) (Everett, 2013; Lazarsfeld et Henry, 1968; McCutcheon, 1987) permet de classifier un ensemble d'individus décrit par des variables qualitatives. Ce modèle est un cas particulier de mélange de distributions. Celui-ci est basé sur les deux hypothèses cruciales suivantes :

- l'ensemble des individus appartenant au même groupe latent ont la même distribution de probabilité ;
- les variables observées dans chaque classe sont indépendantes.

Dans le cas des séquences catégorielles $(\mathbf{z}_i)_{1\leq i\leq n}$ avec $\mathbf{z}_i = (\mathbf{z}_1, \ldots, \mathbf{z}_\mathrm{T})$ où $\mathbf{z}_{it} \in \{1, \ldots, \mathrm{K}\}$, ce modèle considère l'existence d'une variable latente $\mathbf{w} = (w_1, \ldots, w_n)$ et est défini par :

$$\mathrm{P}(\mathbf{z}_i) = \sum_{g=1}^{\mathrm{G}} p_g \prod_{t=1}^{\mathrm{T}} \prod_{k=1}^{\mathrm{K}} z_{itk} \mathrm{P}(z_{it} = k | w_i = g; \boldsymbol{\beta}_g), \tag{3.11}$$

où G est le nombre des classes latentes, p_g est la probabilité a priori d'appartenir à la classe latente g et $z_{itk} = 1$ si $z_{it} = k$ et 0 sinon. Les paramètres du modèle des classes latentes $\boldsymbol{\phi}_g = (p_g, \boldsymbol{\beta}_g)$ sont estimés à l'aide de la méthode du maximum de vraisemblance. La log-vraisemblance est donnée par :

$$\mathrm{L}(\boldsymbol{\phi}) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{\mathrm{G}} p_g \prod_{t=1}^{\mathrm{T}} \prod_{k=1}^{\mathrm{K}} \mathrm{P}(z_{it} = k | w_i = g; \boldsymbol{\beta}_g) \right]. \tag{3.12}$$

Ce critère ne peut pas être maximisé directement. Pour ce faire, l'algorithme EM permet d'estimer les paramètres du modèle en alternant les étapes suivantes jusqu'à la convergence :

- Étape E : consiste à calculer les probabilités a posteriori d'appartenance de chaque séquence à la classe latente g :

$$\tau_{ig}^{(q)} = \frac{p_g \prod_{t=1}^{\mathrm{T}} \prod_{k=1}^{\mathrm{K}} z_{itk} \mathrm{P}(z_{it} = k | w_i = g; \boldsymbol{\beta}_g)}{\sum_{g'=1}^{\mathrm{G}} p_{g'} \prod_{t=1}^{\mathrm{T}} \prod_{k=1}^{\mathrm{K}} z_{itk} \mathrm{P}(z_{it} = k | w_i = g'; \boldsymbol{\beta}_{g'})}; \tag{3.13}$$

- Étape M : consiste à mettre à jour les paramètres du modèle en utilisant les probabilités a posteriori calculées dans l'étape E :

$$p_g^{(q+1)} \quad \leftarrow \quad \frac{\sum_{i=1}^{n} \tau_{ig}^{(q)}}{n}, \tag{3.14}$$

  où les paramètres $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g\ell})_{1\leq \ell \leq \mathrm{K}}$ peuvent être estimés à l'aide de l'estimateur du maximum de vraisemblance.

Une fois les paramètres du modèle estimés, les séquences peuvent être affectées aux classes latentes. Ce modèle permet de regrouper les séquences en se basant sur la proportion des modalités observées sur toute la période et ne tient pas compte de la dépendance temporelle des variables observées. Une extension de ce modèle permettant d'intégrer les variables quantitatives est proposée par Agresti et Kateri (2011).

Cette méthode est appliqué dans de plusieurs domaines (Cho et collab., 2013; Lanza et Collins, 2008; Li et collab., 2016; Martin et collab., 1996; Velicer et collab., 1996).

Application dans le domaine des Smart Grids

Les habitudes de consommation des usagers d'électricité ont été analysées par Kwac et collab. (2014), Wang et collab. (2016), Melzi et collab. (2017) et Labeeuw et Deconinck (2013). Kwac et collab. (2014) ont proposé une méthodologie pour regroupement des consommateurs d'électricité. Cette méthode utilise dans un premier temps un dictionnaire pour encoder les profils de consommation d'électricité. En se basant sur ce dernier, les consommateurs sont regroupés en fonction d'une mesure de variabilité (*entropy of shape*). Cette étude a conduit à une segmentation des consommateurs en fonction de leur variabilité (stable, modéré et variable).

Melzi et collab. (2017) ont proposé une approche de classification non supervisée pour l'extraction des profils-types de consommation à partir des données collectées via les compteurs intelligents. Cette approche est basée sur un mélange de gaussiennes dont les paramètres varient

en fonction du type de jour (semaine, samedi et dimanche). Ainsi, pour chaque composante du mélange, le modèle fournit trois profils de consommation associés à chaque type de jour. Les résultats obtenus mettent en évidence la différence entre les profils de consommation regroupés. Les catégories socioéconomiques des usagers sont utilisées a posteriori pour interpréter les résultats obtenus.

Les modèles de mélange gaussiens ont également été utilisés pour regrouper la demande en eau dans un réseau d'eau intelligent (McKenna et collab., 2014). La méthode utilisée a permis d'identifier les principaux motifs de la demande au fil du temps et de différencier des profils d'usage des consommateurs résidentiels et des consommateurs d'une zone commerciale. Aksela et Aksela (2010) ont proposé également une classification des consommateurs d'eau en se basant sur leurs consommations hebdomadaires moyennes. La méthode *k-means* a été utilisée dans ce cadre. Ainsi, quatre groupes de consommateurs sont obtenus; chaque groupe étant caractérisé par son propre niveau de consommation hebdomadaire.

Wang et collab. (2016) ont proposé une méthodologie constituée de deux étapes : les données de consommation sont d'abord discrétisées à l'aide d'une approximation symbolique agrégée (SAX) (Lin et collab., 2003), ce qui permet de réduire la taille des données, et dans un second temps, un modèle de Markov est utilisé pour modéliser le comportement évolutif des consommateurs. Le partitionnement des consommateurs est effectué en utilisant une technique de classification basée sur la densité des observations. Cette méthode opère sur une matrice de similarité calculée en utilisant la distance Kullback-Liebler entre chaque paire de matrices de transition. L'influence des variables exogènes telles que la température et les évènements calendaires n'a pas été étudiée.

3.1.2 Méthodes de prévision des séries temporelles

Dans cette section, nous décrivons d'abord deux approches de prévision propres aux données catégorielles, à savoir la méthode de régression logistique multinomiale et les modèles de Markov non homogènes. Ensuite, nous présentons un état de l'art sur l'application de ces méthodes dans le domaine des Smart Grids.

Les méthodes de prévision opèrent généralement en deux étapes qui sont les suivantes : La première étape est dédiée à l'estimation des paramètres sur une base de données appelée *base d'apprentissage*. Une fois que les paramètres sont estimés, la deuxième étape consiste à effectuer les prévisions sur une base de données non observée appelée *base de test*. Les performances des méthodes de prévision sont évaluées sur la base de test.

Modèle de régression logistique multinomiale

Le modèle de régression logistique multinomiale fait partie des méthodes d'apprentissage supervisé qui s'appuie sur la distribution multinomiale. Cela permet de modéliser la probabilité d'appartenance d'une observation z_t à une catégorie $k \in \{1,\dots,\mathrm{K}\}$. Dans le cadre d'une séquence temporelle $(\mathbf{z}_t)_{1\leq t\leq \mathrm{T}}$, z_t correspond à une observation à l'instant t. Cette probabilité s'écrit comme suit :

$$\pi_k(z_t) = \mathrm{P}(z_t = k|\mathbf{u}_t;\boldsymbol{\theta}), \tag{3.15}$$

où $(\mathbf{u}_t) \in \mathcal{R}^m$ désigne le vecteur des descripteurs associé qui est de dimension m, $\boldsymbol{\theta}$ est le vecteur des paramètres du modèle à estimer et nous avons $\sum_{k=1}^{\mathrm{K}} \pi_k(z_{it}) = 1$.

Le schéma 3.3 montre la représentation graphique d'un modèle de régression logistique pour une séquence catégorielle $\mathbf{z}$ et les vecteurs de descripteurs associés $\mathbf{u}$. Dans ce schéma, l'observation $\mathbf{z}_{\mathrm{T}+1}$ (affiché par un carré en gris) correspond à une observation non observée.

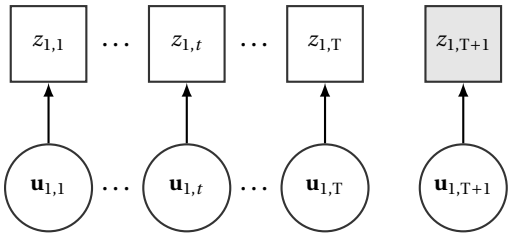


FIGURE 3.3 – Représentation graphique du modèle de régression logistique pour une séquences $\mathbf{z}$ et les vecteurs de descripteurs associés $\mathbf{u}$

Le vecteur des paramètres $\boldsymbol{\theta}$ peut être estimé à l'aide de l'algorithme *iteratively reweighted least squares* (IRLS) (HOLLAND et WELSCH, 1977). Une fois les paramètres estimés, les observations non observées peuvent être prédites à l'aide de la formule suivante :

$$\hat{z}_{iT+1} = \arg\max_k P(z_{iT+1} = k|\mathbf{u}_{iT+1}; \hat{\boldsymbol{\theta}}). \tag{3.16}$$

Modèle de Markov non homogène

Les chaînes de Markov constituent un cadre adapté pour la prévision des séries temporelles (EDWARDS et collab., 2012; FOKIANOS et KEDEM, 1998; KUMAR et JAIN, 2010). Les modèles de Markov permettent de modéliser la dynamique de l'évolution des états au sein d'une séquence. Les modèles de Markov non-homogènes (décrit dans les sections précédentes) permettent d'inclure les variables d'entrée dans le modèle. Le schéma 3.4 montre la représentation graphique d'un modèle de Markov non-homogène pour une séquence $(z_1, \ldots, z_{T+1})$ et ses vecteurs de descripteurs associés $(\mathbf{u}_1, \ldots, \mathbf{u}_{T+1})$.

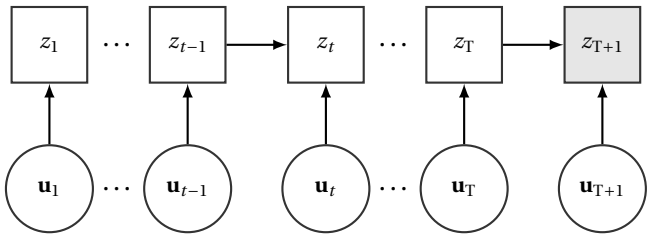


FIGURE 3.4 – Modèle de Markov non homogène

Ce modèle permet de prédire un état futur z_{T+1} en s'appuyant sur les paramètres estimés du modèle $\hat{\boldsymbol{\theta}}$ jusqu'à l'instant T et sur le vecteur des descripteurs associé $\mathbf{u}_{iT+1}$. Pour un état z_{iT+1}, la prévision est fournie à l'aide de la formule suivante :

$$\hat{z}_{iT+1} = \arg\max_k P(z_{T+1} = k|z_{iT}, \mathbf{u}_{iT+1}; \hat{\boldsymbol{\theta}}). \tag{3.17}$$

Application dans le domaine des Smart Grids

Des modèles à base de réseaux de neurones sont généralement utilisés pour prédire le niveau de consommation d'eau (ALTUNKAYNAK et NIGUSSIE, 2017; JAIN et collab., 2001; WALKER et collab., 2015). ALTUNKAYNAK et NIGUSSIE (2017) ont proposé une méthodologie basée sur les réseaux de neurones pour prédire le niveau de consommation à l'échelle mensuel. Cette méthode est constituée de deux étapes : le prétraitement des données en utilisant un modèle de saisonnalité multiplicative (MSA) et la prévision de la tendance en utilisant un réseau de neurones Perceptron multicouches (MLP). Cette méthodologie a également été appliquée sur les données de

consommation brutes. Les résultats obtenus montrent une meilleure performance de la méthodologie proposée grâce à l'étape du prétraitement qui permet de tenir compte de la périodicité des données de consommation.

Les performances des réseaux de neurones pour la prévision à court terme de la consommation d'eau sont comparées par rapport aux méthodes basées sur la régression linéaire (BOUGADIS et collab., 2005). L'influence des variables climatiques (niveau de précipitation, température) est également prise en compte. Les résultats obtenus montre de meilleures performances des réseaux de neurones pour la prévision da la demande. Il a également été montré que la consommation hebdomadaire était corrélée avec le niveau de précipitation. Dans une étude similaire, les réseaux de neurones artificielles sont utilisés par ADAMOWSKI (2008) pour la prévision des pics journaliers de la consommation d'eau.

CANDELIERI et ARCHETTI (2014) ont proposé une méthode de prévision à court terme opérant en deux étapes : les motifs de la demande sont d'abord identifiés en utilisant une méthode de classification, puis les consommations horaires à l'échelle d'un jour sont prédites en utilisant la méthode SVM. Un modèle SVM est estimé par classe et pour chaque heure de la journée. Cette approche a été validée en utilisant les données de consommation issues du réseau d'eau de Milan, pour lesquelles les vraies valeurs de consommation étaient disponibles.

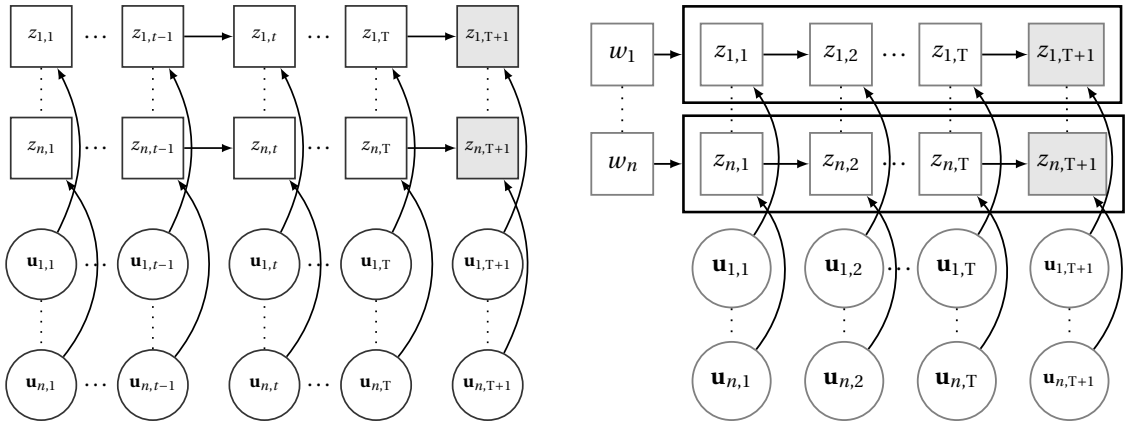
Les chaînes de Markov sont utilisées par GAGLIARDI et collab. (2017) pour la prévision à court terme de la demande en eau. Deux approches sont comparées : l'une basée sur les modèles de Markov homogènes et l'autre sur les modèles de Markov non homogènes. Ce travail a eu pour objectif de prédire le niveau de consommation à l'échelle d'un jour (24 heures). Le niveau de consommation est la seule variable d'entrée pour le modèle de Markov non homogène. les résultats obtenus ont montré de meilleures performances du modèle de Markov homogène pour une prévision à court terme.

3.2 Méthodologie proposée basée sur la modélisation markovienne non homogène

L'approche proposée pour le regroupement des séquences catégorielles notées $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ avec $\mathbf{z}_i = (z_{it})_{1 \leq t \leq \mathrm{T}}$, est basée sur un mélange de chaînes de Markov non-homogènes. Chaque séquence $\mathbf{z}_i$ est constituée de T valeurs observées durant une période commune à l'ensemble des séquences. Cette approche permet de classifier les séquences catégorielles en tenant compte de leur lien avec des variables exogènes, chaque classe étant représentée par sa propre dynamique markovienne non homogène. Ainsi, le modèle proposé suppose que chaque séquence $\mathbf{z}_i$ est issue d'un des G groupes en exploitant une variable latente $w_{i(i=1,\ldots,n)}$ associée à chacune des séquences. Par ailleurs, dans chaque groupe g, les séries temporelles sont modélisées par un modèle de Markov non-homogène d'ordre un et en utilisant les variables d'entrées $\mathbf{u}$ (BENGIO, 1999; BENGIO et FRASCONI, 1996). Ce modèle respecte la propriété de Markov suivante :

$$\mathrm{P}(z_{it} \mid z_{i1}, \ldots, z_{it-1}, \mathbf{u}_{i1}, \ldots, \mathbf{u}_{it}, w_i) = \mathrm{P}(z_{it} \mid z_{it-1}, \mathbf{u}_{it}, w_i). \tag{3.18}$$

Selon cette propriété, l'état actuel z_{it} dépend seulement de l'état à l'instant précédent z_{it-1} et du vecteur des variables d'entrées $\mathbf{u}_{it}$ au sein d'un groupe indiqué par w_i. Les figures 3.5a et 3.5b donnent respectivement la représentation graphique du modèle intégrant une seule dynamique markovienne appelé *Joint Non-homogeneous Markov Model* (JNMM) et celle correspondant au mélange de chaînes de Markov non-homogènes appelé *Mixture of Joint Non-homogeneous Markov models* (MixJNMM).



(a) Modèle JNMM (b) Modèle MixJNMM

FIGURE 3.5 – Représentation graphique probabiliste du modèle présentant une seule dynamique markovienne (appelé JNMM) (a) et d'un mélange de chaînes de Markov non-homogènes (appelé MixJNMM) (b). Dans ces schémas, $z_{i,t}$ désigne l'état de la séquence i à l'instant t, $\mathbf{u}$ est le vecteur des variables exogènes et w est la variable latente qui gère le partitionnement des séquences.

En examinant les schémas présentés dans la figure ci-dessus, la principale différence entre les deux modèles tient au fait que le modèle de mélange (voir figure 3.5b) inclut des variables aléatoires supplémentaires $w_1, \ldots, w_n$ qui gèrent le partitionnement des compteurs. Dans ce modèle, chaque état z_{it} représente un comportement adopté par une entité i issue du groupe w_i et à l'instant t. Ce modèle est défini par le mélange de densités conditionnelles

$$f(\mathbf{z}_i|\mathbf{u}_i;\boldsymbol{\phi}) = \sum_{g=1}^{\mathrm{G}} p_g f_g(\mathbf{z}_i|\mathbf{u}_i;\boldsymbol{\phi}_g), \tag{3.19}$$

où G désigne le nombre de classes, p_g est la probabilité a priori de la classe g satisfaisant $\sum_{g=1}^{\mathrm{G}} p_g = 1$, $\boldsymbol{\phi}_g$ constitue l'ensemble des paramètres du modèle de Markov à estimer pour chaque classe g et $f_g(.)$ indique la distribution associée à un modèle de Markov pour une série temporelle issue de la classe g :

$$f_g(\mathbf{z}_i|\mathbf{u}_i;\boldsymbol{\phi}_g) = \mathrm{P}(z_{i1} \mid w_i = g, \mathbf{u}_{i1};\boldsymbol{\phi}_g) \prod_{t=2}^{\mathrm{T}} \mathrm{P}(z_{it} \mid w_i = g, z_{it-1}, \mathbf{u}_{it};\boldsymbol{\phi}_g). \tag{3.20}$$

Le modèle logistique multinomial est utilisé pour formaliser la dépendance par rapport aux variables de contexte $\mathbf{u}_{it}$ comme suit :

$$\mathrm{P}(z_{i1} = \ell \mid w_i = g, \mathbf{u}_{i1}) = \pi_{g\ell}(\mathbf{u}_{i1};\boldsymbol{\alpha}_g) = \frac{\exp(\boldsymbol{\alpha}_{g\ell}^{\top}\mathbf{u}_{i1})}{\sum_{\ell'=1}^{\mathrm{K}} \exp(\boldsymbol{\alpha}_{g\ell'}^{\top}\mathbf{u}_{i1})}, \tag{3.21}$$

$$\mathrm{P}(z_{it} = k \mid z_{it-1} = \ell, w_i = g, \mathbf{u}_{it}) = \pi_{g\ell k}(\mathbf{u}_{it};\boldsymbol{\beta}_{g\ell}) = \frac{\exp(\boldsymbol{\beta}_{g\ell k}^{\top}\mathbf{u}_{it})}{\sum_{k'=1}^{\mathrm{K}} \exp(\boldsymbol{\beta}_{g\ell k'}^{\top}\mathbf{u}_{it})}, \tag{3.22}$$

où $\pi_{g\ell}$ désigne les probabilités initiales du modèle de Markov pour une classe g qui dépendent des variables d'entrée et dont le vecteur des paramètres est $\boldsymbol{\alpha}_g = (\boldsymbol{\alpha}_{g\ell})_{\ell=1,\ldots,\mathrm{K}}$, et $\pi_{g\ell k}$ désigne les probabilités de transition du modèle de Markov pour une classe g qui dépendent également des variables d'entrée et dont le vecteur des paramètres est $\boldsymbol{\beta}_{g\ell} = (\boldsymbol{\beta}_{g\ell k})_{\ell=1,\ldots,\mathrm{K}}$. La section suivante présente l'algorithme développé pour l'estimation des paramètres du modèle.

3.2.1 Algorithme d'estimation des paramètres

Pour estimer la valeur optimale des paramètres du modèle, la méthode habituelle consiste à maximiser le critère de log-vraisemblance suivant :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}) &= \log \mathrm{P}(\mathbf{z}_1,\ldots,\mathbf{z}_\mathrm{T} \mid \mathbf{u};\boldsymbol{\phi}) = \log \prod_{i=1}^{n} \mathrm{P}(\mathbf{z}_i \mid \mathbf{u}_i;\boldsymbol{\phi}) \\ &= \sum_{i=1}^{n} \log \sum_{g=1}^{\mathrm{G}} p_g \big[\mathrm{P}(z_{i1} \mid w_i = g, \mathbf{u}_{i1};\boldsymbol{\phi}_g) \prod_{t=2}^{\mathrm{T}} \mathrm{P}(z_{it} \mid z_{it-1}, w_i = g, \mathbf{u}_{it};\boldsymbol{\phi}_g) \big],\end{aligned} \tag{3.23}$$

où $\boldsymbol{\phi} = (p_g, \boldsymbol{\alpha}_g, \boldsymbol{\beta}_g)$ est l'ensemble des paramètres du modèle, avec $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g\ell k})_{\ell k}$. Les termes $\mathrm{P}(z_{i1} \mid w_i = g, \mathbf{u}_{i1};\boldsymbol{\phi}_g)$ et $\mathrm{P}(z_{it} \mid z_{it-1}, w_i = g, \mathbf{u}_{it};\boldsymbol{\phi}_g)$ sont obtenus à partir des équations (3.21) et (3.22). Ce critère ne peut pas être maximisé directement. Pour ce faire, on peut utiliser l'algorithme EM (Expectation Maximization) (Dempster et collab., 1977). L'algorithme EM est un algorithme d'estimation itératif qui permet de maximiser la vraisemblance par des mises à jour successives des paramètres du modèle. Les paramètres du modèle $(\boldsymbol{\alpha}_g^{(q+1)}, \boldsymbol{\beta}_{g\ell}^{(q+1)})$ peuvent être calculés en maximisant la quantité :

$$\mathcal{L}_1(\boldsymbol{\alpha}_g) + \sum_{\ell=1}^{\mathrm{K}} \mathcal{L}_{2,\ell}(\boldsymbol{\beta}_{g\ell}), \tag{3.24}$$

où

$$\mathcal{L}_1(\boldsymbol{\alpha}_g^{(q+1)}) = \sum_{\ell=1}^{\mathrm{K}} \sum_{i=1}^{n} \tau_g^{(q)}(\mathbf{z}_i) z_{i1\ell} \log \pi_{g\ell}(\mathbf{u}_{i1};\boldsymbol{\alpha}_g^{(q+1)}), \tag{3.25}$$

et

$$\mathcal{L}_{2,\ell}(\boldsymbol{\beta}_{g\ell}^{(q+1)}) = \sum_{t=2}^{\mathrm{T}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n} \tau_g^{(q)}(\mathbf{z}_i) z_{it-1\ell}\, z_{itk} \log \pi_{g\ell k}(\mathbf{u}_{it};\boldsymbol{\beta}_{g\ell}^{(q+1)}), \tag{3.26}$$

avec

$$\tau_g^{(q)}(\mathbf{z}_i) \leftarrow \frac{p_g^{(q)} f_g(\mathbf{z}_i;\boldsymbol{\phi}_g^{(q)})}{\sum_{g'=1}^{\mathrm{G}} p_{g'}^{(q)} f_{g'}(\mathbf{z}_i;\boldsymbol{\phi}_{g'}^{(q)})} \tag{3.27}$$

étant la probabilité a posteriori d'appartenance d'une séquence $\mathbf{z}_i$ au groupe g.Les variables indicatrices suivantes permettent de spécifier des transitions au sein de chaque séquence i :

- $z_{i1\ell} = 1$ si $z_{i1} = \ell$, sinon $z_{i1\ell} = 0$;
- $z_{itk} = 1$ si $z_{it} = k$, sinon $z_{itk} = 0$ otherwise ;
- $z_{itk} z_{i(t-1)\ell} = 1$ si $z_{it} = k$ et $z_{i(t-1)} = \ell$ simultanément, sinon $z_{itk} z_{i(t-1)\ell} = 0$.

Compte tenu de la grande taille des données étudiées, dans cette thèse, l'algorithme CEM (Classification EM) proposé par Celeux et Govaert (1992) a été adopté pour estimation des paramètres du modèle de Markov ainsi que la variable aléatoire w qui encode la partition. Ce dernier converge plus rapidement que l'algorithme EM. Il consiste à maximiser la vraisemblance complétée suivante :

$$\begin{aligned}\mathrm{P}(\mathbf{z},\mathbf{w} \mid \mathbf{u};\boldsymbol{\phi}) &= \mathrm{P}(\mathbf{w})\,\mathrm{P}(\mathbf{z} \mid \mathbf{w},\mathbf{u};\boldsymbol{\phi}) \\ &= \prod_{i=1}^{n} \mathrm{P}(w_i)\,\mathrm{P}(\mathbf{z}_i \mid w_i,\mathbf{u}_i;\boldsymbol{\phi}_{w_i}).\end{aligned} \tag{3.28}$$

En développant l'équation (3.28), on obtient la log-vraisemblance complétée suivante :

$$\begin{aligned}\mathcal{CL}(\boldsymbol{\phi}, \mathbf{w}) &= \log \mathrm{P}(\mathbf{z}, \mathbf{w} \mid \mathbf{u}; \boldsymbol{\phi}) \\ &= \sum_{g=1}^{\mathrm{G}} \left[\sum_{i=1}^{n} w_{ig} \log p_g + \sum_{i=1}^{n} \sum_{\ell=1}^{\mathrm{K}} w_{ig} z_{i1\ell} \log \pi_{g\ell}(\mathbf{u}_{i1}; \boldsymbol{\alpha}_g) \right. \\ &+ \left. \sum_{i=1}^{n} \sum_{t=2}^{\mathrm{T}} \sum_{k=1}^{\mathrm{K}} \sum_{\ell=1}^{\mathrm{K}} w_{ig} z_{itk} z_{i(t-1)\ell} \log \pi_{g\ell k}(\mathbf{u}_{it}; \boldsymbol{\beta}_{g\ell}) \right],\end{aligned} \tag{3.29}$$

où $w_{ig} = 1$ si $w_i = g$, et $w_{ig} = 0$ sinon. De même que pour l'algorithme EM, les paramètres du modèle peuvent être estimés en maximisant la quantité :

$$\mathcal{CL}_1(\boldsymbol{\alpha}_g) + \sum_{\ell=1}^{\mathrm{K}} \mathcal{CL}_{2,\ell}(\boldsymbol{\beta}_{g\ell}). \tag{3.30}$$

Afin de renforcer la robustesse dans l'estimation des paramètres du modèle de Markov, un terme de régularisation a été ajouté à la fonction de vraisemblance (3.29) :

$$\underset{(\boldsymbol{\phi}, \mathbf{w})}{\arg\max} \left(\mathcal{CL}(\boldsymbol{\phi}, \mathbf{w}) - \frac{\xi}{2} \|\boldsymbol{\phi}\|_2^2 \right), \tag{3.31}$$

où $\|.\|_2$ désigne la norme L_2 et ξ est un hyper-paramètre qui contrôle l'importance du terme de régularisation. Dans le cadre de l'inférence bayésienne (Gauvain et Lee, 1994), ce problème d'estimation des paramètres peut être résolu à l'aide de la méthode du maximum a posteriori (MAP).

Par conséquent, le problème (3.31) consiste à résoudre $\mathrm{G} \times (\mathrm{K}+1)$ problèmes de maximisation distincts :

$$\begin{aligned}&\underset{\boldsymbol{\beta}_{g\ell}}{\arg\max} \left[\mathcal{CL}_{2,\ell}(\boldsymbol{\beta}_{g\ell}) - \frac{\xi}{2} \|\boldsymbol{\beta}_{g\ell}\|_2^2 \right], \quad \forall \; g, \ell \\ &\underset{\boldsymbol{\alpha}_g}{\arg\max} \left[\mathcal{CL}_1(\boldsymbol{\alpha}_g) - \frac{\xi}{2} \|\boldsymbol{\alpha}_g\|_2^2 \right], \quad \forall \; g = 1, \ldots, \mathrm{G}.\end{aligned} \tag{3.32}$$

En d'autres termes, l'étape M de l'algorithme CEM consiste à résoudre $\mathrm{G} \times (\mathrm{K}+1)$ problèmes de régression logistique pour estimer les paramètres du modèle de Markov non-homogène au sein de chaque groupe des compteurs. Une variante de l'algorithme de Newton (Roos et collab., 1998), connue dans cette situation sous le nom « Iteratively Reweighted Least Squares » (IRLS) (Holland et Welsch, 1977), est utilisée pour l'estimation de ces paramètres. Elle garantit une convergence rapide vers une solution optimale. Pour ce faire, cet algorithme nécessite des expressions du vecteur gradient et de la matrice hessienne. Ces derniers sont fournis dans l'annexe A.1. Le pseudo-code 3 présente l'algorithme CEM dédié à l'estimation des paramètres du modèle.

Algorithme 3 : Algorithme CEM-MixJNMM

Entrées : n séquences $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$, vecteur des variables d'entrées associées $(\mathbf{u}_1,\ldots,\mathbf{u}_n)$, nombre de classes G, paramètre initial $\boldsymbol{\phi}^{(0)} = (\boldsymbol{\alpha}_g^{(0)}, \boldsymbol{\beta}_{g\ell}^{(0)})$, partition initiale $\mathbf{w}^{(0)}$

$q \leftarrow 0$;

répéter

Étape E : calcul des probabilités à posteriori :

$$\tau_g^{(q)}(\mathbf{z}_i) \leftarrow \frac{p_g^{(q)} f_g(\mathbf{z}_i;\boldsymbol{\phi}_g^{(q)})}{\sum_{g'=1}^{\mathrm{G}} p_{g'}^{(q)} f_{g'}(\mathbf{z}_i;\boldsymbol{\phi}_{g'}^{(q)})}$$

Étape C : calcul de la partition $\mathbf{w}^{(q+1)}$:

$$w_i^{(q+1)} = \underset{1\leq g\leq \mathrm{G}}{\arg\max}\, \tau_g^{(q)}(\mathbf{z}_i)$$

Étape M : mise à jour des paramètres :

$$p_g^{(q+1)} \quad \leftarrow \quad (1/n)\sum_{i=1}^{n} w_{ig}^{(q)}$$

$i \leftarrow 0$;

répéter

$$\boldsymbol{\alpha}_g^{(i+1)} \quad \leftarrow \quad \boldsymbol{\alpha}_g^{(i)} - \left(\frac{\partial^2 \mathcal{CL}_1(\boldsymbol{\alpha}_g^{(i)})}{\partial\, \boldsymbol{\alpha}_{gk} \partial\, \boldsymbol{\alpha}'_{g\ell}}\right)^{-1} \times \frac{\partial\, \mathcal{CL}_1(\boldsymbol{\alpha}_g^{(i)})}{\partial\, \boldsymbol{\alpha}_{g\ell}}$$

$$\boldsymbol{\beta}_{g\ell}^{(i+1)} \quad \leftarrow \quad \boldsymbol{\beta}_{g\ell}^{(i)} - \left(\frac{\partial^2 \mathcal{CL}_{2,\ell}(\boldsymbol{\beta}_{g\ell}^{(i)})}{\partial\, \boldsymbol{\beta}_{gk\ell} \partial\, \boldsymbol{\beta}'_{gh\ell}}\right)^{-1} \times \frac{\partial\, \mathcal{CL}_{2,\ell}(\boldsymbol{\beta}_{g\ell}^{(i)})}{\partial\, \boldsymbol{\beta}_{gk\ell}}$$

$i \leftarrow i+1$;

jusqu'à ce que l'algorithme IRLS converge;

$q \leftarrow q+1$;

$\boldsymbol{\phi}_g^{(q)} = (p_g^{(q)}, \boldsymbol{\alpha}_g^{(i)}, \boldsymbol{\beta}_{g\ell}^{(i)})_{1\leq g\leq \mathrm{G}}$

jusqu'à ce que la vraisemblance converge;

Sorties : $\widehat{\boldsymbol{\phi}} = \boldsymbol{\phi}_g^{(q)}$; $1 \leq g \leq \mathrm{G}$ et $\widehat{\mathbf{w}} = \mathbf{w}^{(q)}$

3.2.2 Choix des paramètres du modèle

Pour la méthode proposée, le meilleur modèle est défini comme le modèle avec un nombre optimal de groupes des séries chronologiques G. Pour sélectionner le nombre de composantes G du mélange, compte tenu de l'algorithme d'estimation adopté (CEM), le critère *Integrated Classification Likelihood* (ICL) (BIERNACKI et collab., 2000) est utilisé. Ce critère est essentiellement le même que le critère d'information Bayésien (BIC) pénalisé par l'entropie $-\sum_{i,g} \tau_g(\mathbf{z}_i)\log\tau_g(\mathbf{z}_i)$:

$$\mathrm{ICL}(\mathrm{G}) = \mathcal{CL}(\hat{\boldsymbol{\phi}}) - \frac{\vartheta(\mathrm{G})}{2}\log(n\times \mathrm{T}), \tag{3.33}$$

où $\hat{\boldsymbol{\phi}}$ est l'estimation de maximum de vraisemblance pour le vecteur des paramètres $\boldsymbol{\phi}$ et $\vartheta(\mathrm{G})$ est le nombre de paramètres du modèle qui est donné par :

$$\vartheta(\mathrm{G}) = \mathrm{G}\times\left[\underbrace{(\mathrm{K}-1)m}_{\boldsymbol{\alpha}_g} + \underbrace{\mathrm{K}\times(\mathrm{K}-1)m}_{(\boldsymbol{\beta}_{g\ell})_{\ell=1,\ldots,\mathrm{K}}}\right] + \underbrace{\mathrm{G}-1}_{(p_g)_{g=1,\ldots,\mathrm{G}}}, \tag{3.34}$$

où K désigne le nombre d'états (journaliers ou hebdomadaires) et m est la dimension du vecteur des variables d'entrée $\mathbf{u}_{it}$. Ce critère est calculé pour différents nombres de classes. Finalement, le modèle correspondant à la valeur optimale du critère est sélectionné.

3.2.3 Variables explicatives contextuelles

Les variables explicatives (continues ou nominales) formant les vecteurs $\mathbf{u}_{it}$ sont utilisées (a priori) conjointement aux séquences catégorielles pour :

- enrichir les modèles en ajoutant du contexte,
- renforcer leurs capacités prédictives.

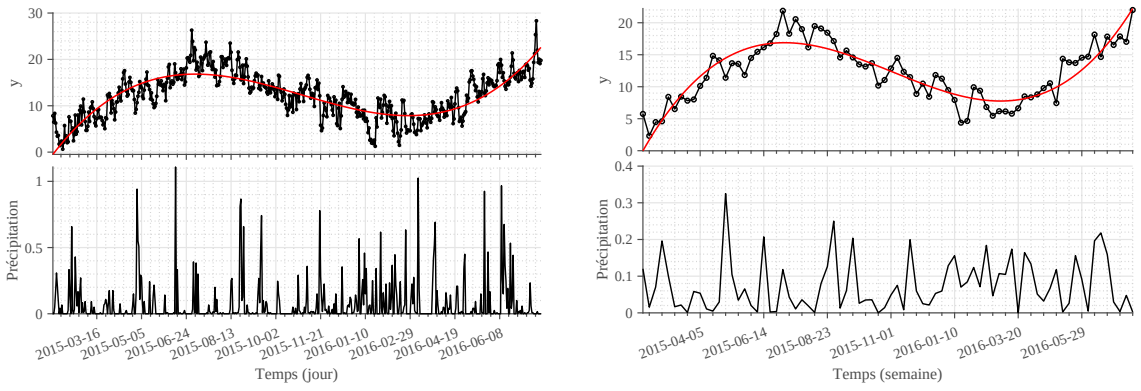
Dans le cas de séquences d'habitudes de consommation issues d'un réseau d'eau, les variables ci-dessous ont été exploitées :

- Température : variable continue qui indique la température moyenne (exprimée en degrés centigrades) de la journée ou de la semaine mesurée par une station météorologique située à Paris ;
- Précipitation : variable continue qui indique la précipitation moyenne (en millimètres) de la journée ou de la semaine mesurée par une station météorologique située à Paris ;
- Évènements calendaires : variable indicative (binaire) qui vaut 1 durant les périodes de vacances scolaires ou pour un jour férié ;
- Saisonnalité : un ensemble de variables continues présentées par des termes trigonométriques qui introduisent le caractère saisonnier des habitudes de consommation au modèle. Elles sont définies comme suit :

$$\mathrm{C}_j(t) = \cos\left(\frac{2\pi j t}{m}\right) \quad \mathrm{S}_j(t) = \sin\left(\frac{2\pi j t}{m}\right). \tag{3.35}$$

  Dans le cas de données à pas de temps journalier, nous avons considéré $m = 7$ (saisonnalité hebdomadaire), et $q \leq \lceil \frac{m}{2} \rceil$ est le nombre requis de termes trigonométriques, où $\lceil \cdot \rceil$ indique la partie entière.

La figure 3.6 montre la représentation graphique des facteurs climatiques. Ces derniers sont mesurés initialement à une fréquence journalière (une valeur par jour). Pour l'analyse du comportement hebdomadaire, ces mesures sont moyennées par semaine. Sur ces graphiques, la tendance de la variable température est affichée en utilisant une courbe rouge. Concernant la variable « Évènements calendaires », nous la présentons en même temps que l'interprétation des résultats de classification.



(a) Cas de profils journaliers

(b) Cas de profils hebdomadaires

FIGURE 3.6 – Représentation graphique des variables météorologiques

3.2.4 Étude expérimentale

Dans cette partie, nous présentons une étude menée d'abord sur des données simulées en vue d'évaluer les performances des méthodes mentionnées pour la classification des séquences catégorielles. Dans un second temps, la méthode proposée est appliquée sur les séquences réelles de consommation issues d'un réseau d'eau potable. Pour les deux cas d'étude, le choix des paramètres des modèles utilisés est justifié à l'aide des méthodes statistiques.

3.2.4.1 Évaluation en termes de classification de séquences

Afin d'évaluer les performances de la méthode proposée en termes de qualité de classification, plusieurs méthodes de l'état de l'art sont évaluées. Celles-ci sont décrites ci-dessous. La représentation graphique de chaque méthode est également rappelée. Dans ces graphiques, des états (catégories) et des variables latentes gérant le partitionnement des séquences sont représentés par les carrés, des variables d'entrée sont indiquées par les cercles et les flèches désignent les relations de dépendance. Les carrés en gris représentent les états futurs qui sont à estimer (voir la section 3.3.1).

- *K-means*. Les séquences (compteurs) sont regroupées en appliquant la méthode des K-moyennes sur p composantes principales (obtenues en appliquant la méthode ACP sur les données non discrétisées) présentant une inertie cumulée d'au moins 90% ;
- *MixMM*. Cette méthode présente un cas spécifique du modèle de mélange proposé où la dynamique de transition au sein des groupes est modélisée à l'aide d'un modèle de Markov homogène (voir figure 3.7). Ainsi, une seule matrice de transition résume l'évolution des états de l'ensemble des séquences pour chaque groupe.

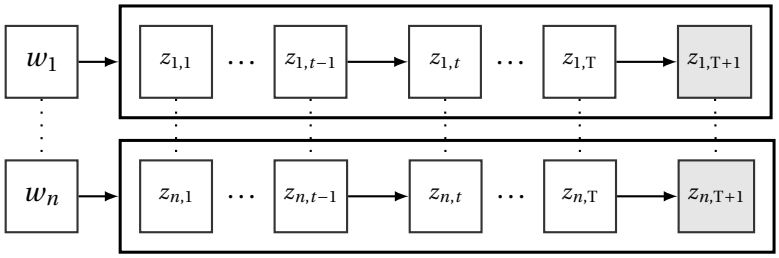


FIGURE 3.7 – Représentation graphique du mélange de modèles de Markov homogènes

- *MixLR*. Cette méthode est une autre variante du modèle de mélange proposé, où l'évolution des états au sein des groupes est modélisée par un modèle de régression logistique (voir figure 3.8). Ce modèle n'intègre pas de dépendance markovienne. Il est défini par le mélange de lois

$$f(\mathbf{z}_i|\mathbf{u}_i;\boldsymbol{\phi}) = \sum_{g=1}^{\mathrm{G}} p_g f_g(\mathbf{z}_i|\mathbf{u}_i;\boldsymbol{\phi}_g), \tag{3.36}$$

où $f_g(\cdot)$ indique la distribution conditionnelle associée aux séries catégorielles issues de la classe g :

$$f_g(\mathbf{z}_i|\mathbf{u}_i;\boldsymbol{\phi}_g) = \prod_{t=1}^{\mathrm{T}} \mathrm{P}(z_{it} \mid w_i = g, \mathbf{u}_{it};\boldsymbol{\phi}_g) \ ; \tag{3.37}$$

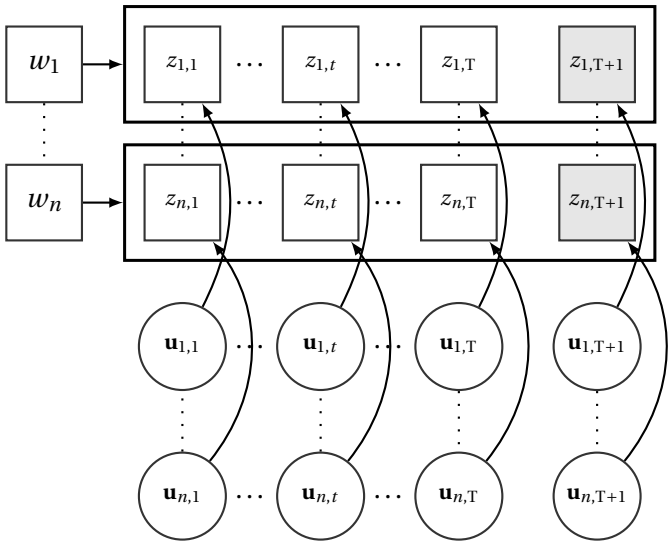


FIGURE 3.8 – Représentation graphique d'un mélange de modèles de régression logistique

- *MixJNMM*. Le modèle de mélange proposé.

3.2.4.2 Critères d'évaluation

La performance des méthodes de classification est évaluée en termes des critères suivants :

- Information Mutuelle Normalisée (NMI) : critère basé sur l'entropie mesurant la dépendance statistique entre deux variables qualitatives, qui sont : les classes estimées et les vraies classes. Ce critère normalisé permet d'effectuer une comparaison entre les différents résultats de classification ;
- Indice de Rand Ajusté (ARI) (HUBERT et ARABIE, 1985) : mesure vérifiant l'accord entre deux partitions ;
- Accuracy : taux des observations (séquences catégorielles) correctement regroupées qui est calculé à partir de la matrice de confusion comme suit :

$$\text{ACC} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}, \tag{3.38}$$

où VP, VN, FP et FN désignent respectivement le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs.

3.2.4.3 Données simulées

Pour comparer les performances des modèles décrits dans la section 3.2.4.1 en termes de qualité de classification, plusieurs bases de données sont générées en utilisant les simulations de type Monte Carlo. Pour cela, nous nous sommes basés sur un jeu de données catégorielles dont les groupes de séquences sont identifiés à l'aide de l'algorithme CEM-MixJNMM. Ce jeu de données (voir figure 3.9) est constitué de 1 526 séquences présentant l'évolution d'états journaliers (10 états) sur une période de 595 jours. Les périodes de vacances scolaires sont indiquées par des flèches sur cette figure. La procédure adoptée pour la simulation, ainsi que l'évaluation des méthodes de classification sur les bases conçues sont abordées dans la suite.

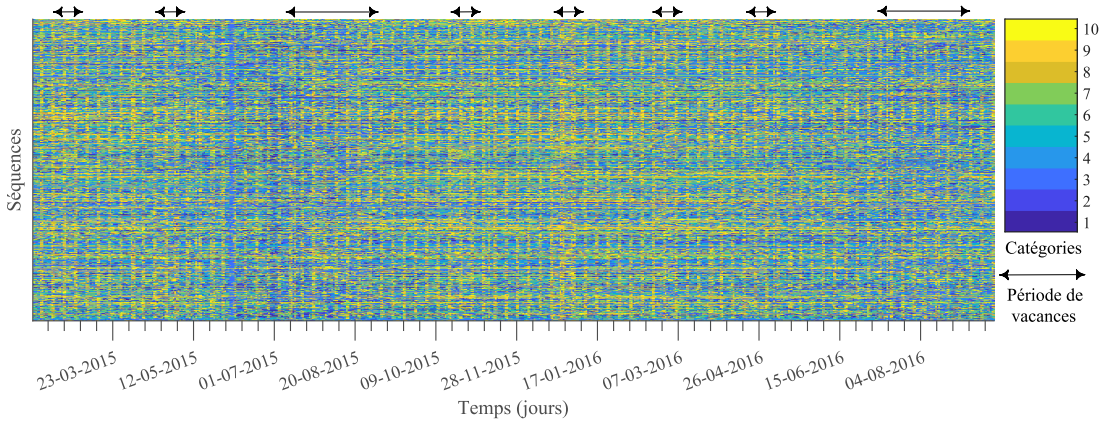


FIGURE 3.9 – Séquences catégorielles représentant l'évolution d'habitudes de consommation de 1,526 compteurs durant 595 jours. Les doubles flèches au-dessus de la figure indiquent les périodes de vacances scolaires.

Procédure adoptée pour les simulations

Pour générer les données, des simulations de type Monte Carlo sont effectuées. À partir des séquences catégorielles $(\mathbf{z}_i)_{1\leq i\leq n}$ et les groupes associés $(w_i)_{1\leq i\leq n}$, un modèle de Markov non-homogène est d'abord estimé par la méthode du maximum de vraisemblance. Les variables exogènes telles que la consommation, les facteurs climatiques, les évènements calendaires, ainsi que les termes trigonométriques sont également utilisées comme variables d'entrée. Finalement, plusieurs séquences sont générées à partir des modèles estimés et trois bases de données sont conçues (voir le tableau 3.1). Celles-ci se distinguent par le nombre de séquences n, leur longueur T, le nombre de classes G et la similarité temporelle des séquences simulées.

TABLEAU 3.1 – Bases de données simulées pour l'évaluation des méthodes de classification

Base de données	`Simulation 1`	`Simulation 2`	`Simulation 3`
(n,T,G)	(200, 595, 2)	(300, 300, 3)	(200, 595, 2)

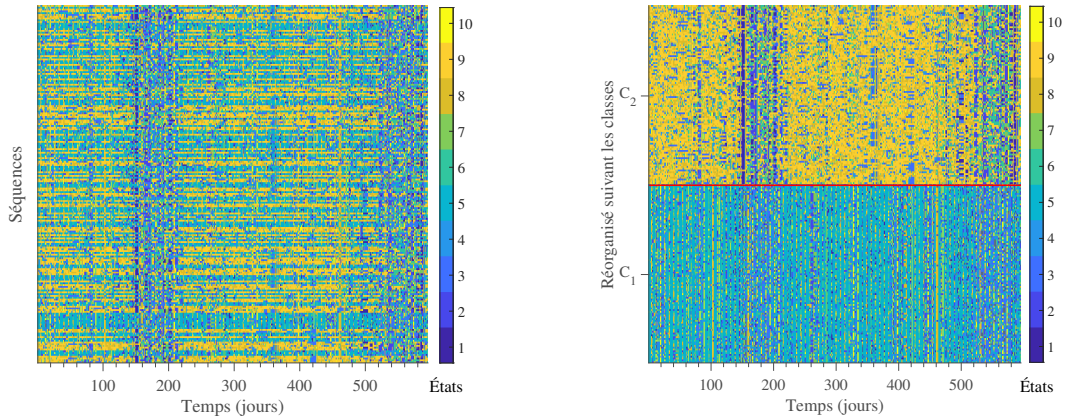
Les séquences simulées et les vraies classes associées sont visualisées conjointement dans les figures 3.10, 3.11 et 3.12. Chaque ligne dans ces graphiques représente une séquence d'états simulée où chaque état est codé par une couleur distincte. Les vraies classes sont séparées par des lignes rouges horizontales. Les méthodes de classification introduites sont appliquées dans un premier temps sur les séquences d'états représentées dans les figures de gauche. Ensuite, les méthodes sont évaluées en utilisant les vraies classes représentées dans les figures de droite.

La première base « Simulation 1 » (voir figure 3.10) est constituée de 200 séquences catégorielles issues de deux dynamiques markoviennes. On peut remarquer une différence significative des dynamiques de transition pour ces deux classes ; ce qui est observable sur toute la période.

Dans un scénario différent (voir figure 3.11), les séquences de la base « Simulation 2 » sont générées à partir d'une même classe. Les paramètres du modèle sont d'abord estimés sur une période de 300 jours et puis séparément sur deux segments de cette période (100 jours et 200 jours). Les paramètres estimés sont notés respectivement $\hat{\boldsymbol{\phi}}_1$, $\hat{\boldsymbol{\phi}}_2$ et $\hat{\boldsymbol{\phi}}_3$. Ainsi, 3 classes de séquences catégorielles sont générées : la première est générée entièrement à partir de $\hat{\boldsymbol{\phi}}_1$, la deuxième est générée en utilisant $\hat{\boldsymbol{\phi}}_2$ et $\hat{\boldsymbol{\phi}}_3$ et la troisième classe est générée en utilisant $\hat{\boldsymbol{\phi}}_3$ et $\hat{\boldsymbol{\phi}}_2$.

La dernière base « Simulation 3 » est constituée de deux groupes de séquences d'états présentant une dynamique de transition très similaires sur une période de 580 jours (voir figure 3.12). Ces séquences sont issues de la même dynamique markovienne estimée à partir de variables d'entrées différentes. Pour la première classe de cette base, seuls les termes trigonométriques (saisonnalité)

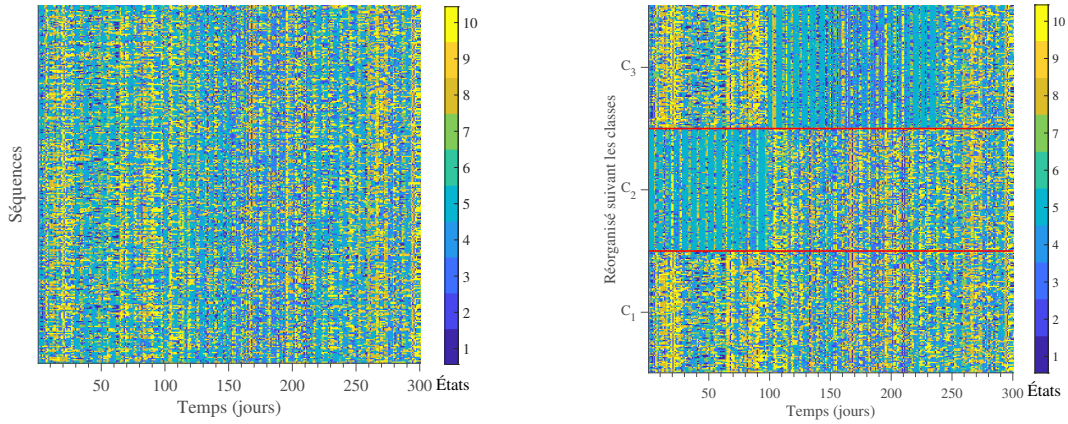
ont été utilisés comme variables d'entrée et pour la deuxième, nous avons ajouté des variables climatiques (température et précipitation).



(a) Base de données « Simulation 1 »

(b) Vraies classes

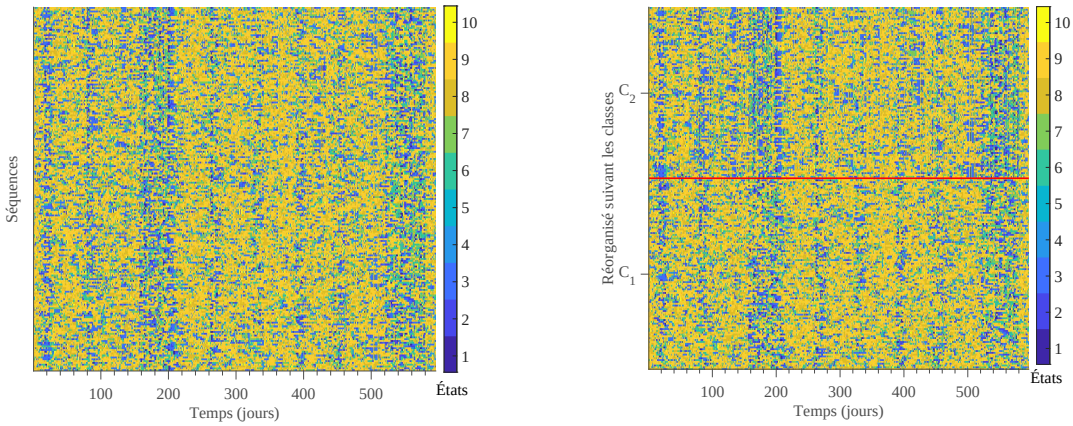
FIGURE 3.10 – Base de données « Simulation 1 » contenant 200 séquences catégorielles sur une période de 595 jours (a) et vraies classes associées (b), les classes sont séparées par la ligne rouge horizontale.



(a) Base de données simulée

(b) Résultat de classification

FIGURE 3.11 – Base de données « Simulation 2 » contenant 300 séquences catégorielles sur une période de 300 jours (a) et vraies classes associées (b), les classes sont séparées par les lignes rouges horizontales.

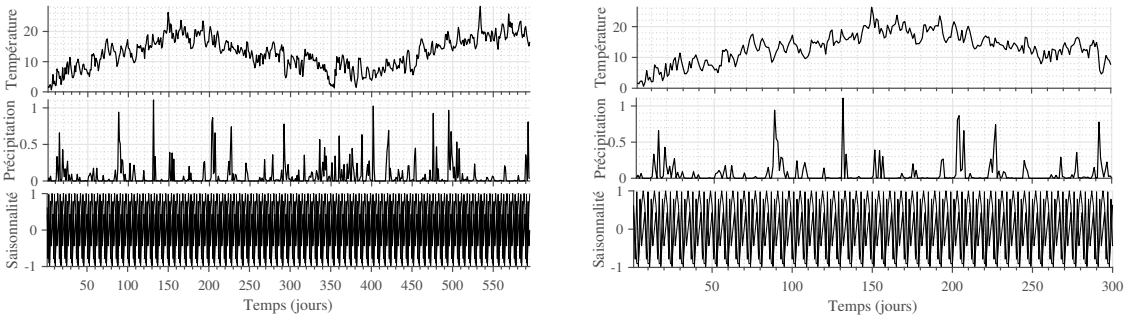


(a) Base de données simulée

(b) Résultat de classification

FIGURE 3.12 – Base de données « Simulation 3 » contenant 200 séquences catégorielles sur une période de 595 jours (a) et vraies classes associées (b), les classes sont séparées par les lignes rouges horizontales.

Les variables exogènes utilisées pour générer les bases de données synthétiques sont présentées dans la figure 3.13. Ces variables sont la température (exprimée en degrés centigrades), les précipitations (en millimètres) et les termes trigonométriques présentant la saisonnalité journalière. Pour les base de données « Simulation 1 » et « Simulation 3 » portant sur une période de 595 jours, les variables représentées dans la figure 3.13a sont utilisées comme variables d'entrée et pour la base « Simulation 2 », les variables utilisées sont celles représentées dans la figure 3.13b.

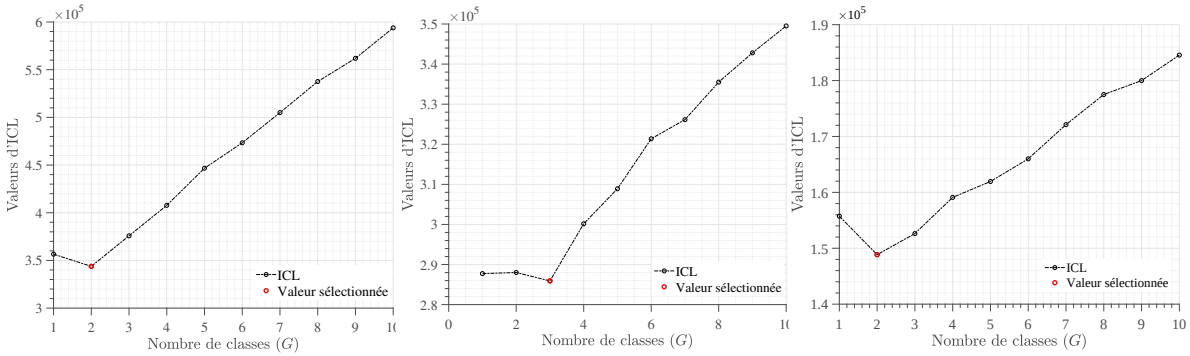


(a) « Simulation 1 » et « Simulation 3 » (b) « Simulation 2 »

FIGURE 3.13 – Variables d'entrées utilisées pour la génération de séquences catégorielles

Choix du nombre de classes

Pour estimer le nombre approprié de classes pour l'algorithme CEM-MixJNMM, nous l'avons exécuté pour différents nombres de classes et nous avons calculé le critère ICL (3.33) pour chacune des bases de données (voir figure 3.14). La valeur sélectionnée de ce critère est indiquée par un cercle rouge sur ces graphiques. On peut remarquer que le critère ICL a permis d'identifier le bon nombre de classes pour les trois bases synthétiques (G = 2 pour les bases « Simulation 1 » et « Simulation 3 » et G = 3 pour la base « Simulation 2 »).



(a) « Simulation 1 » (b) « Simulation 2 » (c) « Simulation 3 »

FIGURE 3.14 – Choix du nombre de classes pour la méthode MixJNMM à l'aide du critère ICL pour les trois bases de données simulées « Simulation 1 » (a), « Simulation 2 » (b) et « Simulation 3 » (c). Le nombre de classes sélectionné pour chacune des bases est indiqué par un cercle rouge.

Comparaison des méthodes de clustering

Cette section vise à comparer les performances des méthodes de classification introduites en utilisant les bases de données synthétiques. Les variables exogènes disponibles sont également considérées comme variables d'entrée pour les méthodes MixLR et MixJNMM. Le tableau 3.2 montre les résultats obtenus à l'aide des critères d'évaluation.

TABLEAU 3.2 – Comparaison des méthodes de classification en termes de différents critères d'évaluation sur les données synthétiques. Les méthodes comparées sont : k-means, mélange de modèles de Markov homogène (MixMM), mélange de modèles de régression logistique (MixLR), mélange de modèles de Markov non-homogène (MixJNMM). Les critères d'évaluations sont : le taux d'observations correctement classifiée (ACC), information mutuelle normalisée (NMI), Indice de rand ajustée (ARI)

Méthodes	**Critères**	`Simulation1`	`Simulation2`	`Simulation3`
k-means	NMI	0,90	0,35	0,18
	ARI	0,94	0,38	0,17
	ACC	0,98	0,73	0,54
MixMM	NMI	1	0,89	0,45
	ARI	1	0,90	0,52
	ACC	1	0,95	0,77
MixLR	NMI	1	0,87	0,56
	ARI	1	0,91	0,60
	ACC	1	0,98	0,83
MixJNMM	NMI	1	1	0,80
	ARI	1	1	0,85
	ACC	1	1	0,96

On peut remarquer les bonnes performances de l'ensemble des méthodes sur la base « Simulation 1 » qui est représentée par deux dynamiques markoviennes très différentes. La bonne performance de la méthode K-means s'explique par le fait qu'une telle différence dans la dynamique de l'évolution des états reflèterait également une différence dans les composantes principales (obtenues par ACP). Concernant la base « Simulation 2 », on peut noter une baisse importante de l'ensemble des critères pour la méthode *k-means*. Les méthodes basées sur le modèle de Markov montrent de bonnes performances et le meilleur résultat est obtenu à l'aide de la méthode proposée. Finalement, on peut remarquer une baisse générale de la performance concernant la base « Simulation 3 ». La méthode proposée fournit le meilleur résultat, ce qui pourrait être du à la capacité de cette méthode à modéliser la dynamique conjointe des données. En revanche, la méthode MixMM montre de très faibles performances, ce qui pourrait être lié au fait que ce modèle ne prend pas en compte des variables exogènes. La méthode MixLR montre une performance légèrement supérieure par rapport à ce dernier grâce à la prise en compte des variables de contexte.

Étude de la sensibilité de la classification par rapport au nombre d'états

Pour analyser la qualité du modèle estimé en fonction du nombre d'états sélectionnés préalablement dans l'étape de discrétisation, une étude de sensibilité de classification est menée sur les données simulées. Ces dernières sont constituées de trois dynamiques markoviennes principales (G = 3) et huit états (K = 8). Nous avons varié rétrospectivement le nombre d'états de 2 à 20, par un pas d'incrément de 2, ce qui mène à la création de dix bases de données catégorielles. Ensuite, les séquences catégorielles issues de chaque base sont regroupées en utilisant la méthode proposée MixJNMM et en fixant le nombre de classes à G = 3. Finalement, le critère de « sensibilité » est calculé entre les résultats de classification obtenus sur chaque base (avec un nombre d'état différent) et les vraies classes issues des données simulées (voir la figure 3.15). Dans notre cas, la sensibilité est estimée par le taux des séquences correctement classifiées. On peut constater que la performance de la méthode proposée ne varie pas de manière significative à partir du nombre d'états K = 8. Ce phénomène s'explique par le fait que le nombre d'états supérieur à 8 pourrait conduire à l'apparition des états redondants.

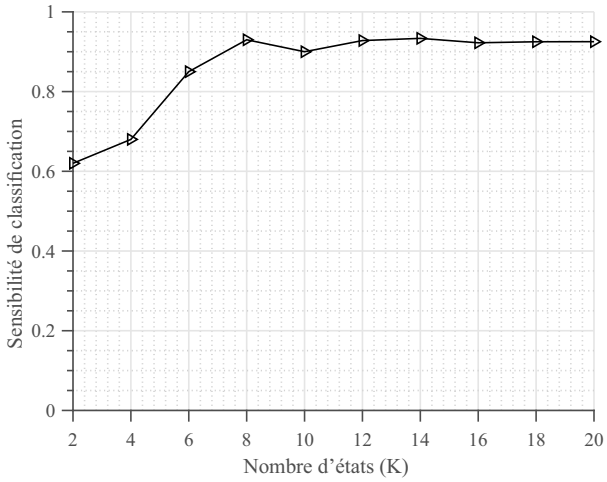


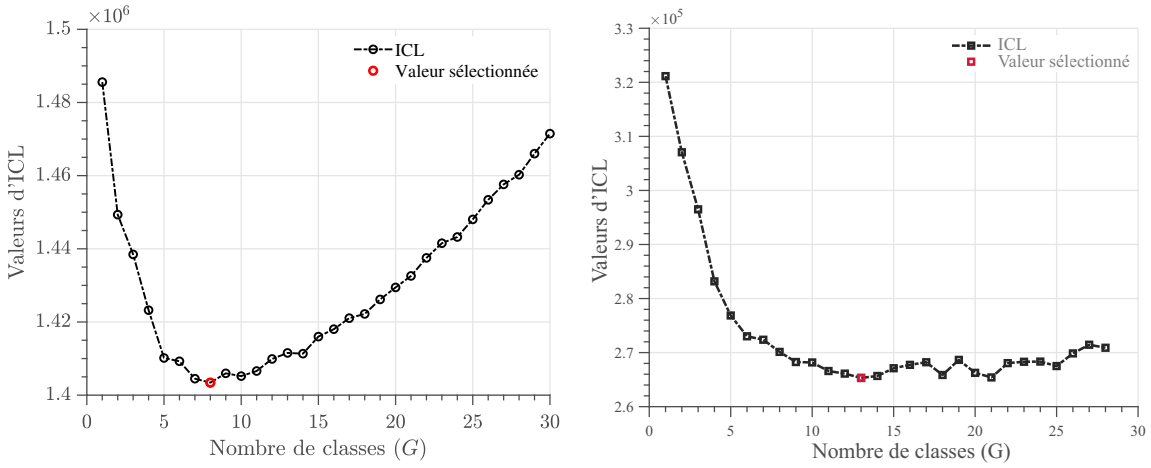
FIGURE 3.15 – Sensibilité de classification en fonction de différents nombres d'états K

3.2.4.4 Données issues du réseau d'eau

Dans cette partie, nous analysons les résultats de l'application de la méthode proposée (MixJNMM) sur les deux bases de données réelles issues d'un réseau d'eau potable (voir figure 3.1). L'objectif est de modéliser la dynamique hétérogène des séquences d'habitudes de consommation et d'interpréter les résultats obtenus à l'aide des outils numériques et graphiques. Les variables exogènes telles que la consommation, les facteurs climatiques et les évènements calendaires sont également utilisées. Cette section commence par décrire le problème lié au choix du nombre de classes.

Choix du nombre de classes

L'objectif visé est de sélectionner le nombre de classes pour les deux bases de données (profils journaliers et profils hebdomadaires). Pour ce faire, le critère ICL (3.33) est calculé pour différents nombres de composantes $G \in \{1,\dots,30\}$ et la valeur minimum de ce critère est retenue. La figure 3.16 montre les résultats associés. La valeur sélectionnée est indiquée en rouge sur les deux graphiques. Pour les bases de séquences d'habitudes journalières et hebdomadaires, nous avons opté respectivement pour les nombres de classes G = 8 et G = 13.



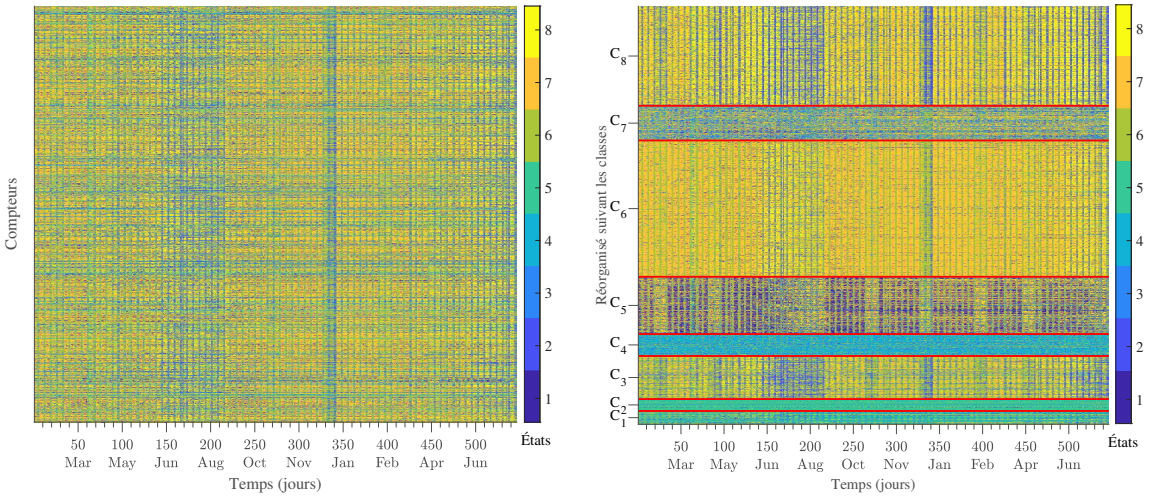
(a) Profils journaliers (G = 8)

(b) Profils hebdomadaires (G = 13)

FIGURE 3.16 – Critère ICL issu de la méthode MixJNMM pour les séquences d'états journaliers (a) et hebdomadaires (b)

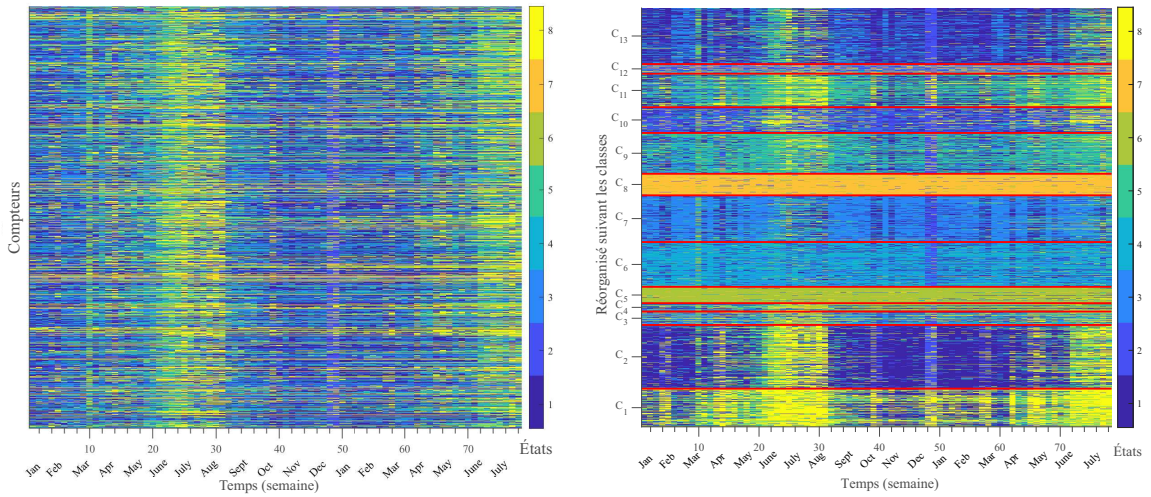
Classification des compteurs

Les partitions obtenues en appliquant l'algorithme MixJNMM avec G = 8 dans le cas de séquences d'états journaliers et avec G = 13 dans le cas de séquences d'états hebdomadaires sont affichées dans les figures 3.17b et 3.18b. Les classes sont séparées par des lignes rouges horizontales. On peut remarquer que les classes sont globalement constituées des habitudes de consommation (états) homogènes dans le temps et la variabilité des habitudes de consommation change d'une classe à l'autre. Par exemple, la classe 2 dans le cas journalier et les classes 5 et 8 dans le cas hebdomadaire sont constituées des habitudes de consommation très homogènes. En revanche, la classe 7 dans le cas journalier et les classes 4 et 11 dans le cas hebdomadaire présentent sont caractérisées par des habitudes de consommation présentant une variabilité importante.



(a) Base de données catégorielles (b) Résultat de classification

FIGURE 3.17 – Classification des séquences d'états journaliers : (a) séquences d'habitudes journalières issues de 2 000 compteurs et (b) 8 classes obtenues (séparées par des lignes rouges horizontales) en utilisant la méthode CEM-MixJNMM



(a) Base de données catégorielles (b) Résultat de classification

FIGURE 3.18 – Classification des séquences d'états hebdomadaires : (a) séquences d'habitudes hebdomadaires issues de 2 000 compteurs et (b) 13 classes obtenues (séparées par des lignes rouges horizontales) en utilisant la méthode CEM-MixJNMM

La matrice de confusion entre les classes issues des séquences d'états hebdomadaires et jour-

naliers a également été calculée (voir tableau 3.3). Cette matrice permet d'établir des correspondances entre certaines classes. Par exemple, les classes 6 et 8 (cas journalier) sont liées aux classes 1, 2, 11 et 13 (cas hebdomadaire). La classe homogène dans le cas des profils journaliers (classe 2) correspond bien à la classe homogène dans le cas des profils hebdomadaires (classe 5). Les classes 3, 4, 5 et 7 (cas journalier) sont respectivement liées aux classes 10, 8, 7 et 6 (cas hebdomadaire). Les résultats de classification obtenues à partir de ces deux discrétisations présentent donc une certaine cohérence.

TABLEAU 3.3 – Matrice de confusion entre les labels des compteurs obtenus par la classification des séquences d'états journaliers et hebdomadaires

		Classes dans le cas journalier							
		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Classes dans le cas hebdomadaire	C_1	0	0	25	0	0	6	0	152
	C_2	1	0	57	0	1	137	0	108
	C_3	5	0	1	0	43	4	4	7
	C_4	12	1	3	2	12	3	4	3
	C_5	24	49	0	2	2	0	1	0
	C_6	1	0	2	0	39	60	111	1
	C_7	1	0	0	0	114	104	2	2
	C_8	18	7	0	76	0	0	2	0
	C_9	0	0	10	0	38	63	10	74
	C_{10}	0	0	96	2	0	4	14	7
	C_{11}	2	0	0	0	15	42	0	101
	C_{12}	0	0	8	23	0	2	12	1
	C_{13}	0	0	4	0	10	227	6	20

Convergence de la méthode proposée

Afin d'illustrer la convergence de l'algorithme CEM-MixJNMM, nous avons affiché la log-vraisemblance calculée à chaque itération de l'algorithme CEM (voir figure 3.19). Dans cette figure, le cercle rouge indique l'itération à partir de laquelle les valeurs de la log-vraisemblance se stabilisent. L'hyper-paramètre ξ qui contrôle l'importance du terme de régularisation est fixé à 10^{-8}. On peut noter la convergence assez rapide de l'algorithme CEM pour l'estimation des paramètres du modèle de mélange.

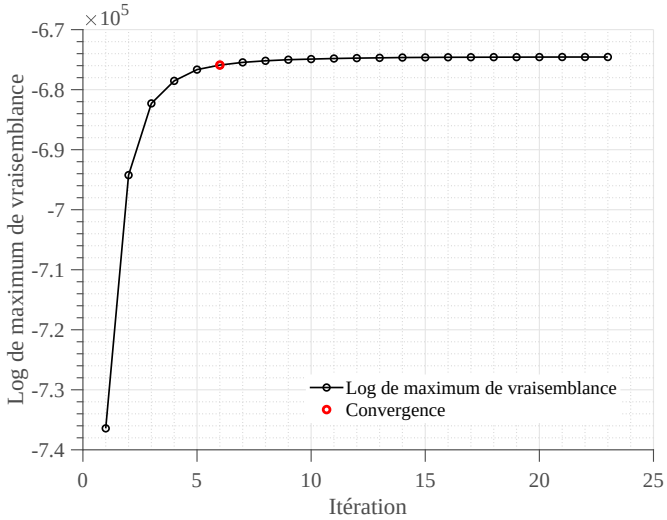


FIGURE 3.19 – Convergence de l'algorithme MixJNMM

Description et interprétation des classes de compteurs

Les figures 3.20 et 3.21 présentent, pour chacune des classes obtenues dans le cas d'états journaliers et hebdomadaires, l'évolution de la proportion des états au fil du temps. Ces graphiques sont très pertinents pour l'interprétation des classes obtenues, dans la mesure où ils permettent de repérer temporellement les périodes de plus forte occurrence des états dans chaque classe. Les périodes de vacances scolaires et de l'été sont encadrées respectivement par des pointillés rouges et cyans dans ces graphiques. Ces périodes peuvent également être identifiées en observant directement les changements dans la proportion des états au fil du temps pour certaines classes. Par exemple dans la figure 3.20, pour la classe 3, on peut observer une hausse importante de l'état 6 (représenté par la couleur vert) durant les périodes de vacances et dans la figure 3.21, pour les classes 1 et 2, on peut observer une hausse importante de l'état 8 (représenté par la couleur jaune) durant l'été. Les section suivantes décrivent les classes identifiées à partir des séquences d'états journaliers et hebdomadaires.

Description des classes (cas journalier)

- *Classe 1 :* cette classe est constituée principalement des états 5 et 7. Elle peut être attribuée aux zones avec les travaux réguliers.
- *Classe 2 :* cette classe est constituée de compteurs avec un comportement très homogène au fil du temps. C'est l'état 5 (consommation constante) qui est le plus présent dans cette classe. Cette classe pourrait correspondre à des sites où des travaux réguliers ont lieu.

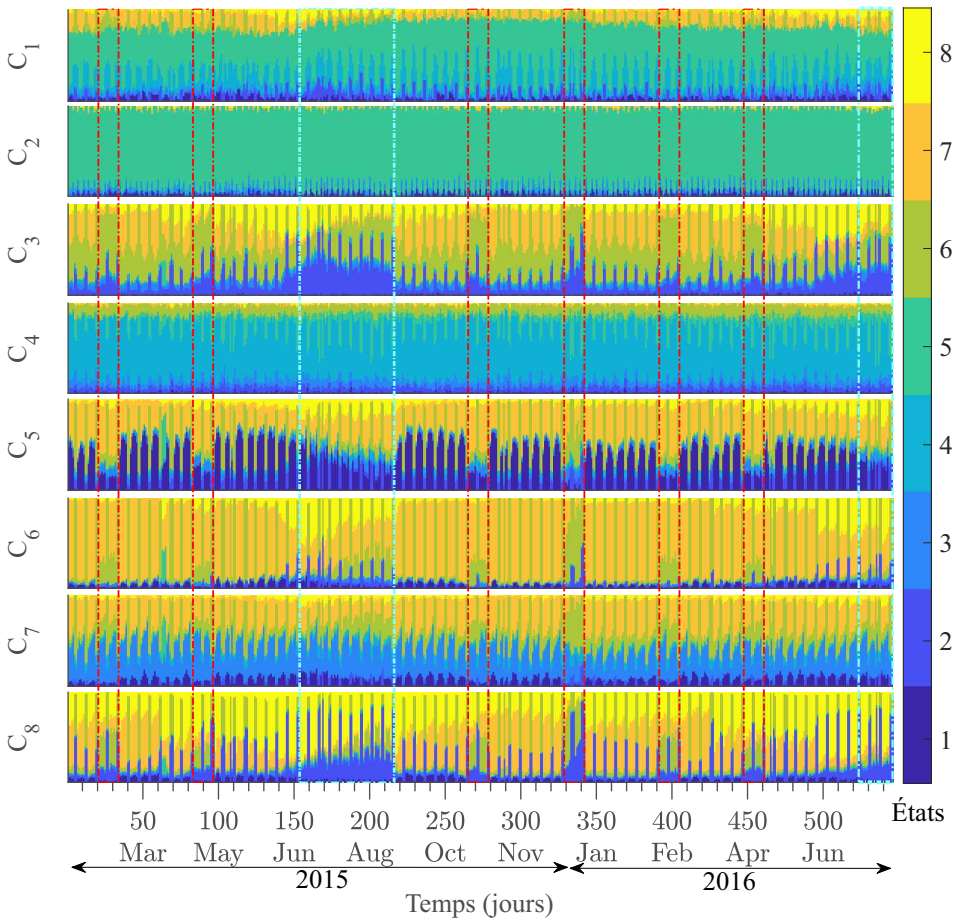


FIGURE 3.20 – Évolution de la proportion des états journaliers selon les classes présentées dans la figure 3.17b. Les périodes de vacances scolaires et d'été sont encadrées par des pointillés rouges et cyans.

- *Classes 3 et 8 :* ces classes représentent aussi un comportement de consommation très varié dans le temps. Plusieurs états sont présents pendant la période de consommation. On peut constater l'apparition de l'état 2 (pic dans l'après-midi et pic du soir) durant les périodes de vacances scolaires et d'été. Elles peuvent être attribuées à des zones actives pendant les périodes de vacances comme la restauration ou les zones touristiques.
- *Classe 4 :* cette classe est dominée par la présence de l'état 4 qui est représenté par un seul pic. Pendant les weekends et les périodes de vacances on remarque l'apparition de l'état 5 qui a une consommation constante. Cette classe peut être associée à une zone d'activité.
- *Classe 5 :* cette classe est constituée globalement des états 1 et 7 qui représentent un profil à deux pics, et avec l'apparition de l'état 6 pendant les weekends et les périodes de vacances. Le pic du matin de l'état 6 se situe bien après les pics du matin des états 1 et 7. Elle pourrait correspondre aux zones comportant des habitats individuels et collectifs.
- *Classe 6 :* cette classe est constituée globalement de l'état 7 pendant l'année scolaire (le pic du matin très tôt et le pic du soir à 20h). Pendant les périodes de vacances scolaires, on peut remarquer l'apparition de l'état 6 (caractérisé par un pic du matin décalé dans le temps). Pendant la période d'été, les états 6 et 8 sont présents (caractérisés par un pic du matin diffus). Cette classe peut être associée aux zones de résidence occupées par des personnes actives avec potentiellement des enfants à charge.
- *Classe 7 :* Cette classe présente un comportement assez varié dans le temps. Les jours ouvrés sont caractérisés par les états 3 et 7. Durant les weekends et les vacances scolaires, on remarque une présence plus forte de l'état 6. On peut associer cette classe aux zones résidentielles et d'activités.

Description des classes (cas hebdomadaire)

- *Classes 1 et 2 :* en regardant l'évolution des états de la classe 1, on peut remarquer que pendant les périodes de vacances, la proportion de l'état 8 (jaune) accroît (caractérisé par un pic du matin plus élevé et plus décalé dans le temps pendant les weekends). La classe 2 respecte à peu près la même dynamique que la classe 1, avec l'apparition plus importante de l'état 3 pendant les vacances scolaires. On peut attribuer cette classe aux zones de résidence occupées par des personnes actives avec potentiellement des enfants à charge.
- *Classe 3 et 4 :* ces deux classes présentent des habitudes de consommation quasi similaires avant la période d'été. Pendant la période d'été, l'état 6 devient l'état dominant dans la classe 4.
- *Classe 5 :* cette classe est constituée principalement de l'état 6 pendant toute la période (état caractérisé par une habitude de consommation symétrique). Elle pourrait correspondre à une zone industrielle avec le déclenchement automatique de machines.
- *Classes 6 et 7 :* ces deux classes représentent à peu près la même évolution d'états. La classe 6 est constituée principalement de l'état 4 alors que la classe 7 est dominée par l'état 3. On peut noter une apparition de l'état 2 pendant les vacances de Noël pour ces deux classes (état caractérisé par un pic du matin décalé et par un pic du soir plus faible). Ces deux classes peuvent être associées à des zones de résidence occupées par des personnes actives.
- *Classe 8 :* cette classe est constituée principalement de l'état 7 (état caractérisé par un seul pic pendant les jours ouvrés et par une consommation constante pendant les weekends). Cette classe pourrait correspondre à une zone d'activité.
- *Classes 9 et 11 :* ces deux classes sont dominées par l'état 5 pendant toute la période et on peut remarquer une apparition des états 1 et 2 pendant les périodes de vacances scolaires. Ces classes peuvent être associées aux zones commerciales avec une activité pendant les périodes de vacances.

- *Classes 10 et 13 :* ces deux classes sont caractérisées principalement par les états 2 et 1 pendant les périodes de vacances scolaires et les périodes hors vacances. On peut observer également une apparition des états 8 et 5 pendant la période d'été.
- *Classe 12 :* cette classe est dominée par la présence des états 3 et 7. Elle peut être associée aux zones résidentielles et d'activités.

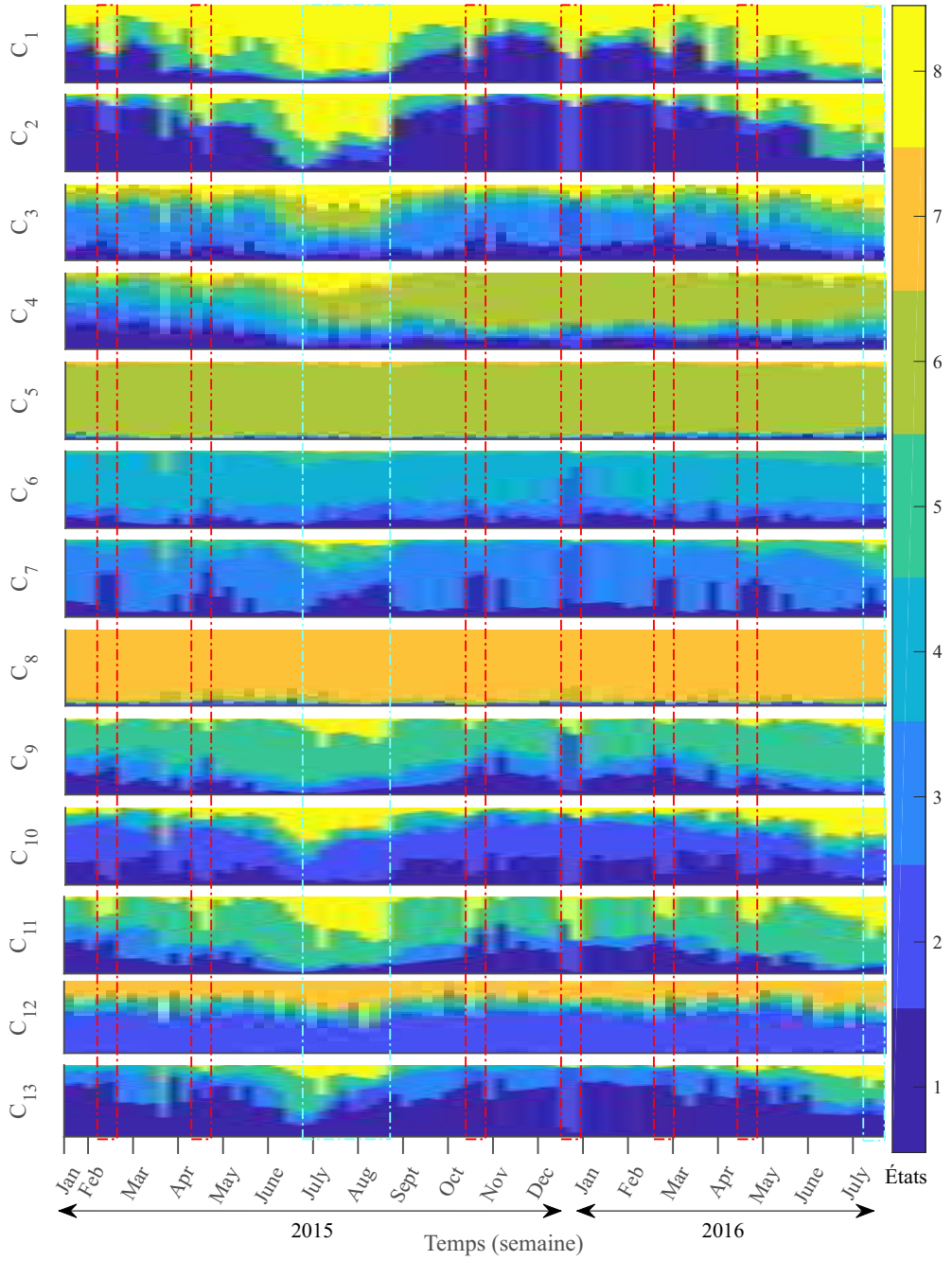


FIGURE 3.21 – Évolution de la proportion des états hebdomadaires selon les classes présentées dans la figure 3.18b. Les périodes de vacances scolaires et d'été sont encadrées par des pointillés rouges et cyans.

Analyse des matrices de transition

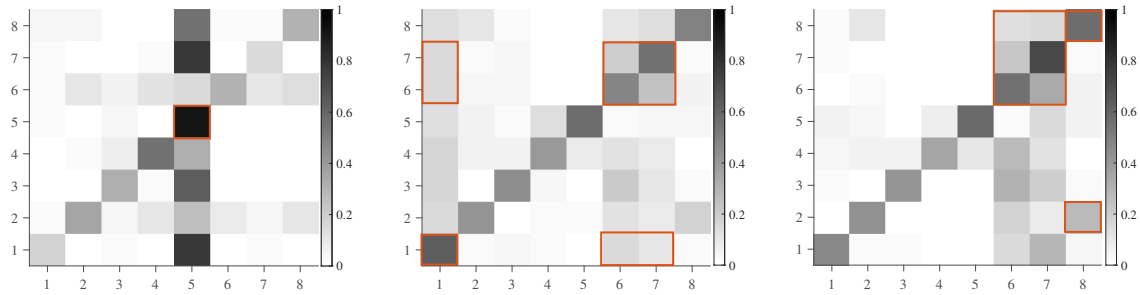
Cette section vise à analyser les transitions entre les habitudes de consommation à l'aide des matrices de transition estimées au sein de chaque classe. Celles-ci permettent en effet de repérer les transitions les plus fréquentes. Pour notre modèle, celles-ci étant dépendantes du temps,

nous les avons moyennées par rapport au temps. On parlera alors de matrices de transitions globales. Ce moyennage peut également être réalisé sur des périodes spécifiques. Dans un premier temps, nous nous focalisons sur les matrices de transition issues de la classification des séquences d'états journaliers. La figure 3.22 montre les matrices de transition globales associées à 3 classes. Les lignes correspondent à l'état $t-1$ et les colonnes correspondent à l'état t. Les transitions prédominantes dans chaque classe sont encadrées par les carrés oranges.

Concernant la classe 2 qui est caractérisée par les habitudes de consommation très homogènes, on observe une probabilité de transition élevée de la plupart des états vers l'état 5 qui est l'état dominant dans cette classe (entouré par un carré orange). Vu le profil de l'état 2, on peut associer les compteurs de cette classe à une zone ayant une consommation très irrégulière (probablement le bruit).

Pour la classe 5, on observe que la plupart des transitions se produisent entre les états 1, 6 et 7; les weekends étant caractérisés par l'état 6 (pic du matin décalé dans le temps) et les jours ouvrés par l'état 1 et 7 respectivement en dehors et pendant les périodes de vacances. On note des transitions de l'état 1 vers l'état 6 lors des vacances de Noël et de l'état 1 vers l'état 7 pendant les autres périodes de vacances.

Finalement, concernant la classe 6 qui est constituée principalement des états 6, 7 et 8, on remarque une probabilité de transition très élevée de l'état 7 vers lui-même (en dehors des périodes de vacances scolaires). En regardant la figure 3.20, on remarque que des transitions entre ces états se produisent notamment au début et à la fin de périodes de vacances. Les compteurs de cette classe pourraient correspondre à des zones de résidence occupées par des personnes actives avec potentiellement des enfants à charge.



(a) Classe 2 (b) Classe 5 (c) Classe 6

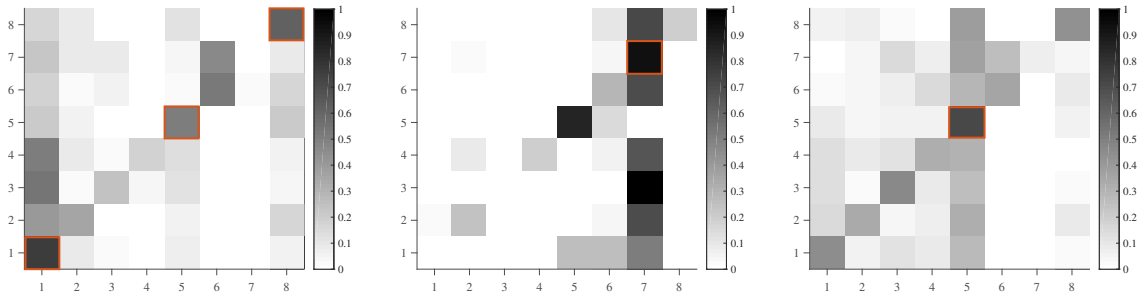
FIGURE 3.22 – Matrices de transition globales correspondant aux 3 classes dans le cas des profils journaliers. Les probabilités élevées sont représentées par une couleur plus foncée. Dans chaque matrice, les transitions prédominantes sont encadrées par des carrés oranges. Les lignes correspondent à l'état $t-1$ et les colonnes correspondent à l'état t.

La figure 3.23 montre les matrices de transition globales associées à 3 classes dans le cas des états hebdomadaires. Pour la classe 2, on observe une probabilité de transition relativement élevée de la plupart des états vers l'état 1 qui est un des états dominants de cette classe. On observe également une probabilité de transition élevée de l'état 8 vers lui-même, qui est un autre état dominant dans cette classe pendant la période d'été. Ainsi, la plupart des consommateurs associés à cette classe changent leur habitude de consommation au début de la période estivale et tendent à garder les mêmes habitudes de consommation jusqu'à la fin des vacances d'été.

Concernant la classe 8 qui est constituée des habitudes de consommation plus homogènes, la plupart des transitions sont vers l'état 7 (seul état dominant dans cette classe) ; la transition de l'état 7 vers lui-même ayant la probabilité la plus élevée. Cette classe pourrait donc être associée aux zones d'activités commerciales ou d'administration.

En analysant la matrice de transition de la classe 9 (classe constituée principalement de l'état 5 pendant les périodes d'été), on remarque une probabilité de transition très élevée de l'état 5 vers

lui-même et une probabilité de transition moyenne de l'ensemble des états vers ce dernier. Les compteurs de cette classe pourraient correspondre aux zones résidentielles avec un membre actif pendant toute l'année.



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

FIGURE 3.23 – Matrices de transition globales correspondant aux 3 classes dans le cas des profils hebdomadaires. Les probabilités élevées sont représentées par une couleur plus foncée. Dans chaque matrice, des états prédominants sont entourés par des carrés oranges. Les lignes correspondent à l'état $t-1$ et les colonnes correspondent à l'état t.

L'analyse des matrices de transition globales a permis de résumer la dynamique des habitudes de consommation au sein des classes de compteurs. Pour caractériser les transitions entre les états, la section suivante est consacrée à l'analyse des facteurs qui pourraient être à l'origine de celles-ci.

Analyse de l'influence des variables exogènes

L'analyse de l'influence des variables de contexte sur les habitudes de consommation a été menée en se basant sur les coefficients estimés du modèle. Pour cela, nous avons extrait les coefficients estimés $\boldsymbol{\beta}$ correspondant aux variables climatiques (température et précipitation) et aux évènements calendaires (vacances scolaires) pour les transitions les plus fréquentes. Cette analyse est menée sur les classes 5, 6 et 8 dans le cas d'états journaliers et sur les classes 2 et 9 dans le cas d'états hebdomadaires. Rappelons dans un premier temps que la probabilité de transition d'un état ℓ vers un état k pour une classe g s'écrit :

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

Les coefficients associés aux classes 6 et 8 dans le cas journalier sont affichés respectivement dans les tableaux 3.4 et 3.5. Dans ces tableaux, chaque valeur correspond à un coefficient estimé du modèle et les valeurs les plus significatives des coefficients sont notées en gras.

En regardant le tableau 3.4, concernant la variable température, on peut noter une valeur relativement élevée (négative) du coefficient pour la transition de l'état 8 vers l'état 7 et une valeur élevée (négative) pour le calendrier. Cela signifie qu'une baisse dans la température, en dehors des vacances scolaires, pourrait conduire à une apparition plus importante de cette transition (transition vers une habitude présentant un pic du matin plus élevé). On peut remarquer une valeur élevée (positive) du coefficient associé à la précipitation pour la transition de l'état 7 vers l'état 6, selon laquelle un niveau de précipitation élevé pourrait accroître cette transition (transition vers une habitude présentant les pics du matin et du soir plus décalés). Finalement, on note une valeur très élevée (positive) du coefficient de la variable calendrier pour la transition de l'état 6 vers lui-même. Cela traduit l'apparition importante de cette transition pendant les périodes de vacances scolaires (notamment pendant les vacances de Noël).

TABLEAU 3.4 – Coefficients estimés des transitions les plus significatives associés aux variables exogènes température, précipitation et calendrier pour la classe 6 dans le cas des états journaliers

Transitions	Coefficient des variables exogènes		
$(s_{t-1} \rightarrow s_t)$	Température	Précipitation	Calendrier
$1 \rightarrow 1$	-0,05	0,52	-0,95
$2 \rightarrow 2$	-0,07	0,10	1,47
$2 \rightarrow 8$	0,01	1,59	1,01
$6 \rightarrow 6$	-0,06	-0,56	**3,33**
$6 \rightarrow 7$	-0,08	-0,68	2,74
$6 \rightarrow 8$	0,07	0,09	3,29
$7 \rightarrow 6$	-0,04	**1,80**	0,11
$7 \rightarrow 7$	-0,06	0,68	-0,19
$8 \rightarrow 6$	-0,03	0,17	0,90
$8 \rightarrow 7$	**-0,17**	-1,30	-0,92
$8 \rightarrow 8$	-0,10	-1,33	0,11

Le tableau 3.5 présente les coefficients de transition pour la classe 8 dans le cas d'états journaliers. On peut noter une valeur relativement élevée (négative) du coefficient de la variable température pour la transition de l'état 8 vers l'état 7 (transition vers une habitude présentant un pic du matin plus élevé). Cette valeur indique qu'une baisse de température pourrait accroître cette transition. Dans la figure 3.20, on peut remarquer la présence de cette transition vers la fin de la période estivale.

TABLEAU 3.5 – Coefficients estimés des transitions les plus significatives associés aux variables exogènes pour la classe 8 dans le cas d'états journaliers

Transitions	Coefficient des variables exogènes		
$(s_{t-1} \rightarrow s_t)$	Température	Précipitation	Calendrier
$2 \rightarrow 2$	-0,01	1,19	0,01
$2 \rightarrow 6$	-0,09	1,12	0,39
$2 \rightarrow 8$	0,01	**1,87**	-0,54
$6 \rightarrow 2$	-0,01	1,08	0,02
$6 \rightarrow 6$	-0,06	0,08	-0,14
$6 \rightarrow 7$	-0,11	-0,64	**-1,39**
$6 \rightarrow 8$	0,01	0,22	-0,64
$7 \rightarrow 6$	-0,07	0,61	-0,69
$7 \rightarrow 7$	-0,10	-0,45	-0,56
$7 \rightarrow 8$	-0,03	-0,44	-0,56
$8 \rightarrow 2$	-0,03	0,71	0,16
$8 \rightarrow 6$	-0,10	0,33	0,21
$8 \rightarrow 7$	**-0,14**	-0,34	-1,14
$8 \rightarrow 8$	-0,05	-0,23	-0,34

D'autres part, le coefficient de la variable calendrier (une valeur élevée négative) confirme l'apparition de cette transition à la fin des vacances scolaires. Concernant la précipitation, on peut remarquer une valeur élevée (positive) du coefficient pour la transition de l'état 2 vers l'état 8 (transition vers une habitude présentant un pic du matin plus tôt et moins élevé). Cela signifie qu'une hausse du niveau de précipitation pourrait correspondre à une apparition importante de cette transition (entrée dans la saison pluvieuse). Finalement, on note également une valeur élevée (négative) du coefficient de la variable calendrier pour la transition de l'état 6 vers l'état 7 (transition vers une habitude présentant un pic du matin plus tôt et un pic du soir plus élevé). Ainsi, les périodes en dehors de vacances scolaires pourraient conduire à cette transition. En re-

gardant le profil de cette classe, on note une proportion plus élevée de l'état 6 pendant les périodes de vacances scolaires (état caractérisé par un pic du matin plus élevé et plus décalé que celui du soir).

Les coefficients associés aux classes 2 et 9 dans le cas d'états hebdomadaires sont affichés respectivement dans les tableaux 3.6 et 3.7.

En examinant le tableau 3.6, on peut noter une valeur élevée (positive) du coefficient pour les transitions de l'état 1 vers les états 5 et 8 (transition vers une habitude présentant un pic du matin décalé). Ainsi, une augmentation de la température pourrait accroître ces transitions (notamment en entrant dans la période estivale). Concernant le coefficient de la variable précipitation, on peut noter une valeur élevée (positive) pour la transition de l'état 8 vers l'état 1 (transition vers une habitude présentant un pic du matin plus tôt et plus élevé), suivant laquelle une hausse dans le niveau de précipitation pourrait accroître cette transition (en début d'automne). On peut également remarquer une valeur élevée (négative) pour les transitions des états 1 et 8 vers l'état 2 (transition vers une habitude présentant un pic du matin plus élevé et plus décalé), selon laquelle un niveau de précipitation faible pourrait accroître l'occurrence de ces transitions. En regardant les coefficients de la variable calendrier, on peut noter des valeurs élevées (négatives) pour les transitions des états 1 et 8 vers l'état 5 (transition vers une habitude présentant un palier entre le pic du matin et celui du soir). Cela signifie qu'en dehors des périodes de vacances, on pourrait remarquer une présence plus importante de cette transition.

TABLEAU 3.6 – Coefficients estimés des transitions les plus significatives associés aux variables exogènes pour la classe 2 dans le cas d'états hebdomadaires

Transitions	Coefficient des variables exogènes		
$(s_{t-1} \rightarrow s_t)$	Température	Précipitation	Calendrier
$1 \rightarrow 1$	0,17	2,13	-0,99
$1 \rightarrow 2$	0,19	**-3,63**	-0,45
$1 \rightarrow 5$	**0,30**	0,68	**-1,40**
$1 \rightarrow 8$	**0,28**	0,41	-0,40
$8 \rightarrow 1$	-0,12	**3,90**	0,52
$8 \rightarrow 2$	-0,06	**-4,77**	-0,02
$8 \rightarrow 5$	-0,01	2,30	**-1,62**

Le tableau 3.7 présente les coefficients des transitions pour la classe 9 dans le cas hebdomadaire. Concernant la variable température, on peut noter une valeur élevée (positive) pour la transition de l'état 8 vers lui-même et une valeur élevée (négative) pour la transition de l'état 5 vers l'état 1 (transition vers une habitude présentant un palier moins marqué entre les pics du matin et

TABLEAU 3.7 – Coefficients estimés des transitions les plus significatives associés aux variables exogènes pour la classe 9 dans le cas d'états hebdomadaires

Transitions	Variables exogènes		
$(s_{t-1} \rightarrow s_t)$	Température	Précipitation	Calendrier
$1 \rightarrow 1$	-0,11	0,46	-0,14
$1 \rightarrow 5$	-0,03	1,75	-0,17
$1 \rightarrow 8$	-0,01	0,44	-0,06
$5 \rightarrow 1$	**-0,17**	1,61	-0,71
$5 \rightarrow 5$	-0,06	2,42	-0,81
$5 \rightarrow 8$	0,03	-1,49	0,13
$8 \rightarrow 1$	-0,03	0,46	-0,57
$8 \rightarrow 5$	0,08	**3,89**	**-1,10**
$8 \rightarrow 8$	**0,14**	-0,69	-0,83

du soir). cela signifie qu'une hausse dans la température pourrait accroître les transitions de l'état 8 vers lui-même. De même, une baisse dans la température pourrait accroître les transitions de l'état 5 vers l'état 1. Concernant la précipitation et le calendrier, on peut noter respectivement une valeur élevée (positive) et une valeur élevée (négative) des coefficients pour la transition de l'état 8 vers l'état 5 (apparition d'un palier entre les deux pics). Cela signifie qu'un niveau de précipitation élevé, surtout en dehors de vacances scolaires, pourrait accroître cette transition.

Dans cette section, les séquences catégorielles ont été regroupées et les groupes obtenus ont été interprétées à l'aide de différents outils numériques et graphiques. Dans la section suivante, nous nous focalisons sur la prévision de futures habitudes de consommation au sein des groupes obtenus.

3.3 Prévision des habitudes de consommation d'eau

Pour prédire les états futurs d'un ensemble de séquences (compteurs), nous avons scindé la base de données en deux parties : une base d'apprentissage qui comporte deux tiers de la période de consommation (année 2015) et une base de test qui est constituée du reste de la période (année 2016). Les paramètres du modèle sont estimés sur la base d'apprentissage. Une fois les paramètres estimés, la formule suivante permet de prédire les futures habitudes de consommation au sein de chaque classe :

$$\begin{aligned} \widehat{z}_{i\mathrm{T}+1} &= \arg\max_k \mathrm{P}(z_{i\mathrm{T}+1} = k \mid z_{i\mathrm{T}}, w_i = g, \mathbf{u}_{i\mathrm{T}+1}) \\ &= \arg\max_k \pi_{w_i z_{i\mathrm{T}} k}(\mathbf{u}_{i\mathrm{T}+1}; \boldsymbol{\beta}_{w_i z_{i\mathrm{T}}}), \end{aligned} \tag{3.39}$$

où $\mathbf{u}_{i\mathrm{T}+1}$ indique le vecteur des variables d'entrée et $\boldsymbol{\beta}_{w_i z_{i\mathrm{T}}}$ indique le vecteur des paramètres estimés sur l'ensemble des séquences de la classe w_i. Le vecteur des variables d'entrée utilisé pour la prévision inclut une ou plusieurs des variables de contexte suivantes :

- T : la température du pas de temps (jour ou semaine) précédent ;
- P : la précipitation du pas de temps précédent ;
- Y : la consommation du pas de temps précédent (un vecteur de taille 24 pour les profils journaliers et de taille 168 pour les profils hebdomadaires) ;
- C : indicateur binaire d'événements calendaires (vacances scolaires, jours fériés) ;
- S : saisonnalité hebdomadaire exprimée en utilisant les termes trigonométriques. Cette variable est utilisée dans le cas de séquences d'états journaliers et cela dans l'ensemble des expérimentations.

3.3.1 Méthodes évaluées

Pour évaluer les performances en termes d'erreurs de prévision de la méthode proposée, celle-ci a été comparée aux méthodes décrites ci-dessous.

- Modèle de Markov homogène (MM) : un cas particulier d'un mélange de modèles de Markov homogène en considérant G = 1. Une matrice de transition résume la dynamique des habitudes de consommation pour l'ensemble de compteurs. En utilisant cette approche, la formule suivante permet de prédire une future habitude de consommation :

$$\begin{aligned} \hat{z}_{i\mathrm{T}+1} &= \arg\max_k \mathrm{P}(z_{i\mathrm{T}+1} = k \mid z_{i\mathrm{T}}) \\ &= \arg\max_k \mathbf{a}_{z_{i\mathrm{T}} k}, \end{aligned} \tag{3.40}$$

  où $\mathbf{a}$ est la matrice de transition estimée sur l'ensemble des compteurs $(\mathbf{z}_{it})_{1\leq i\leq n, 1\leq t\leq \mathrm{T}}$;

- Mélange de modèles de Markov homogènes (MixMM) : il s'agit de la même méthodologie que l'approche proposée mais sans prise en compte des variables de contexte. La formule suivante permet de prédire les habitudes de consommation futures pour un compteur i :

$$\begin{aligned} \hat{z}_{i\mathrm{T}+1} &= \arg\max_k \mathrm{P}(z_{i\mathrm{T}+1} = k \mid z_{i\mathrm{T}}, w_i) \\ &= \arg\max_k \mathbf{a}_{w_i z_{i\mathrm{T}} k}, \end{aligned} \tag{3.41}$$

  où $\mathbf{a}_{w_i}$ est la matrice de transition estimée au sein d'un groupe w_i ;

- La régression logistique (LR) : un cas particulier d'un mélange de régressions logistiques en considérant G = 1. La dynamique globale des habitudes de consommation est estimée à partir d'un seul modèle de régression logistique. Ce modèle prend en compte les variables contextuelles. Les futures habitudes de consommation sont prédites à partir de la formule suivante :

$$\begin{aligned} \hat{z}_{i\mathrm{T}+1} &= \arg\max_k \mathrm{P}(z_{i\mathrm{T}+1} = k \mid \mathbf{u}_{i\mathrm{T}+1}) \\ &= \arg\max_k \pi(\mathbf{u}_{i\mathrm{T}+1}; \boldsymbol{\phi}); \end{aligned} \tag{3.42}$$

- Mélange de régressions logistiques (MixLR) : il s'agit de la même méthodologie que l'approche proposée mais sans la dépendance markovienne. La formule suivante permet de prédire les habitudes de consommations futures pour un compteur i :

$$\begin{aligned} \hat{z}_{i\mathrm{T}+1} &= \arg\max_k \mathrm{P}(z_{i\mathrm{T}+1} = k \mid w_i, \mathbf{u}_{i\mathrm{T}+1}) \\ &= \arg\max_k \pi_{w_i}(\mathbf{u}_{i\mathrm{T}+1}; \boldsymbol{\phi}_{w_i}); \end{aligned} \tag{3.43}$$

- Modèle de Markov non-homogène (JNMM) : un cas particulier du modèle de mélange proposé en considérant G = 1. Il suppose que l'ensemble des compteurs suit la même dynamique markovienne non homogène. En utilisant cette approche, les futures habitudes de consommation sont prédites à l'aide de la formule suivante :

$$\begin{aligned} \widehat{z}_{i\mathrm{T}+1} &= \arg\max_k \mathrm{P}(z_{i\mathrm{T}+1} = k \mid z_{i\mathrm{T}}, \mathbf{u}_{i\mathrm{T}+1}) \\ &= \arg\max_k \pi_{z_{i\mathrm{T}} k}(\mathbf{u}_{i\mathrm{T}+1}; \boldsymbol{\beta}_{z_{i\mathrm{T}}}); \end{aligned} \tag{3.44}$$

- K-means + JNMM : cette méthode est constituée de deux étapes réalisées séparément : la classification des séries initiales (à fréquence horaire) à l'aide de l'algorithme des K-means et la prévision des séries catégorielles au sein de chaque classe en utilisant un modèle de Markov non-homogène ;
- Mélange de modèle de Markov non-homogènes (MixJNMM) : modèle proposé.

Les avantages(+) et les inconvénients (-) des méthodes comparées sont résumés dans le tableau 3.8.

TABLEAU 3.8 – Avantages et inconvénients des méthodes comparées

Méthodes	Avantages	Inconvénients
MM	Prise en compte de la régularité temporelle, complexité algorithmique faible	Pas de prise en compte de variables exogènes
LR	Prise en compte de variables exogènes, complexité algorithmique moyenne	Pas de prise en compte de la régularité temporelle
JNMM	Prise en compte de variables exogènes et de la régularité temporelle	Pas de prise en compte de structure locale, complexité algorithmique moyenne
MixJNMM	Prise en compte d'une structure locale au sein des séquences	Complexité algorithmique élevée

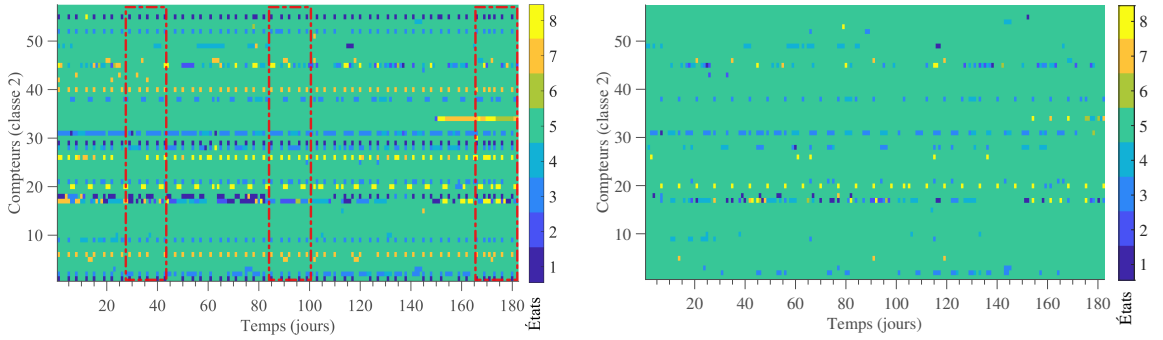
3.3.2 Critères d'évaluation de la qualité des prévisions

Les performances en termes d'erreurs de prévision des méthodes décrites sont comparées en utilisant les critères suivants :

- Indice de Rang Ajusté (ARI) : une mesure vérifiant l'accord entre le résultat de prévision et les vrais labels ;
- Accuracy : taux d'états futurs correctement prédits ;
- Rappel : défini pour chaque catégorie par le taux d'observations correctement prédites parmi toutes les observations de la catégorie ;
- Précision : définie pour chaque catégorie par le taux d'observations correctement prédites parmi les observations affectées à cette catégorie ;
- F-mesure : moyenne harmonique de la précision et du rappel.

3.3.3 Prévision des habitudes de consommation issues du réseau d'eau potable

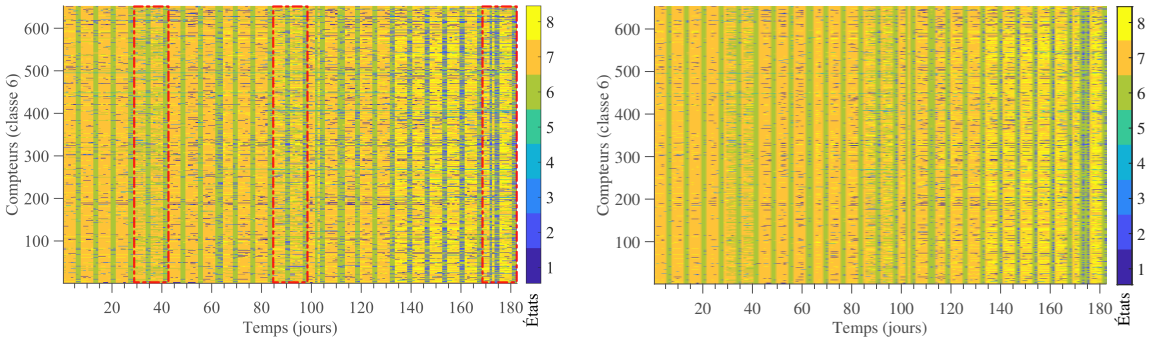
Les figures 3.24 et 3.25 présentent les résultats de prévision à 1 pas de temps obtenus en appliquant la méthode proposée (MixJNMM) sur les séquences issues des classes 2 et 6 dans le cas d'états journaliers. Les périodes de vacances sont encadrées par des pointillés rouges dans les figures de gauche. Ces dernières présentent un extrait de données réelles (base de test) portant sur l'intervalle de prévision (182 jours). La classe 2 est constituée des habitudes de consommation homogènes et les séquences de la classe 6 montrent un comportement plus variable.



(a) Données réelles (b) Prévision

FIGURE 3.24 – Prévision des habitudes journalières de la classe 2. Les périodes de vacances sont encadrées par des pointillés rouges dans le graphique de gauche qui présente un extrait de données réelles (base de test) sur l'intervalle de prévision (182 jours en 2016).

Dans le cas d'une classe constituée principalement des habitudes de consommation homogènes (voir figure 3.24), les prévisions effectuées sont assez précises. On peut remarquer que pen-



(a) Données réelles (b) Prévision

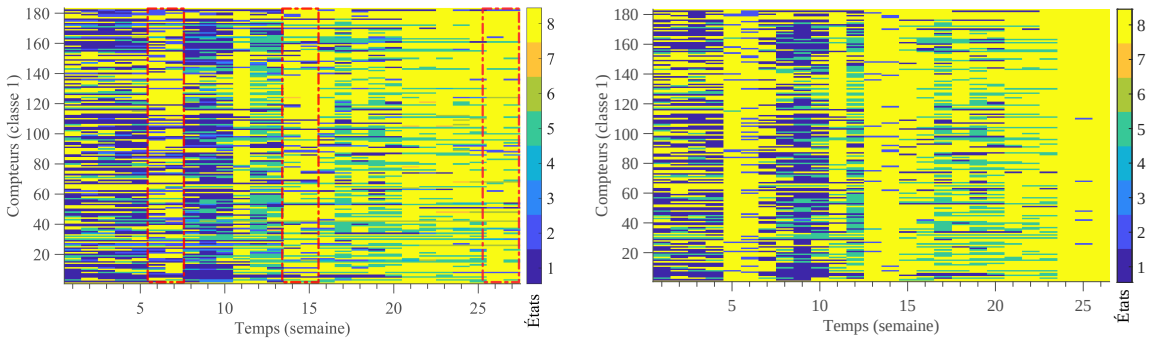
FIGURE 3.25 – Prévision des habitudes journalières de la classe 6. Les périodes de vacances sont encadrées par des pointillés rouges dans le graphique de gauche qui présente un extrait de données réelles (base de test) sur l'intervalle de prévision (182 jours en 2016).

dant les périodes de vacances, les compteurs associés à cette classe montrent les mêmes habitudes de consommation qu'en dehors de ces périodes, ce qui facilite la prévision de leurs habitudes de consommation.

Dans le cas d'une classe non-homogène (voir figure 3.25), on remarque la saisonnalité hebdomadaire pendant les périodes hors vacances (périodes en dehors des pointillés rouges). Les transitions entre les jours ouvrés et les weekends sont correctement prédites en tenant compte de la saisonnalité des habitudes journalières. Les prévisions effectuées sur les séquences de la classe 6 (voir figure 3.25b) sont relativement proches des données réelles (voir figure 3.25a). Cependant, on remarque des prévisions moins précises pendant les périodes de vacances qui est dû à la variabilité importante des habitudes de consommation pendant ces périodes.

Les figures 3.26 et 3.27 présentent les résultats de prévision obtenus en appliquant la méthode proposée sur les séquences des classes 1 et 12 dans le cas d'états hebdomadaires. Rappelons que la classe 1 présente une variabilité moyenne et la classe 12 présente une variabilité élevée de ses habitudes de consommation.

Dans le cas d'une classe constituée d'états moyennement hétérogènes (voir figure 3.26), les prévisions sont assez précises pendant les périodes hors vacances et celles-ci sont moins précises pendant les périodes de vacances scolaires (périodes encadrées par des pointillés rouges). Cela traduit la irrégularité de l'évolution des états pendant ces périodes qui est difficile à prédire.

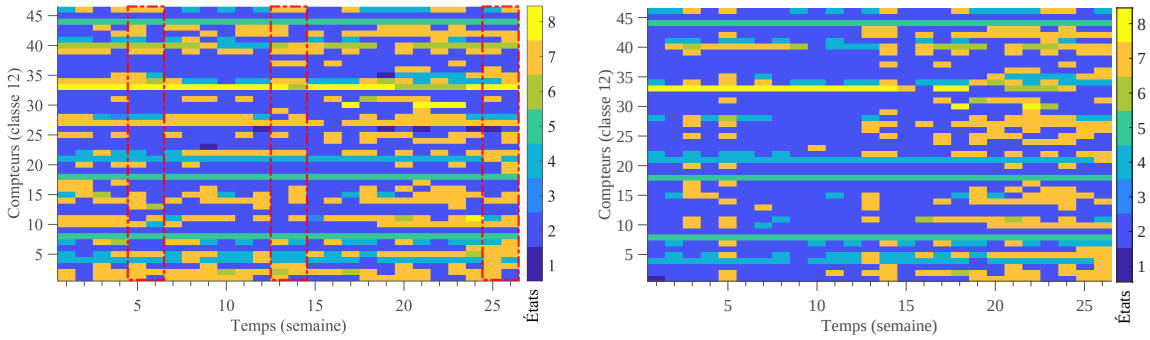


(a) Données réelles (b) Prévision

FIGURE 3.26 – Prévision des habitudes hebdomadaires de la classe 1. Les périodes de vacances sont encadrées par des pointillés rouges dans le graphique de gauche qui présente un extrait de données réelles (base de test) sur l'intervalle de prévision (26 semaines en 2016).

Concernant la classe 12 (voir figure 3.27) qui est constituée des séquences présentant une évolution d'états très hétérogène, on remarque des précisions moins précises. On note des prévisions

plus précises pour les séquences constituées d'états homogènes et les prévisions moins précises pour les séquences constituées d'états hétérogènes au sein de cette classe.



(a) Données réelles (b) Prévision

FIGURE 3.27 – Prévision des habitudes hebdomadaires de la classe 12. Les périodes de vacances sont encadrées par des pointillés rouges dans la figure de gauche qui présente un extrait de données réelles (base de test) sur l'intervalle de prévision (26 semaines en 2016).

3.3.4 Comparaison des méthodes

Les tableaux 3.9 et 3.10 montrent les résultats obtenus à l'aide des méthodes mentionnées et en utilisant différentes combinaisons des variables d'entrée. Dans ces tableaux, les performances de chaque méthodes sont également représentées par des barres dont la longueur est proportionnelle à la valeur numérique de l'erreur. Tous les critères d'évaluation utilisés prennent des valeurs dans l'intervalle $[0,1]$. Les meilleures performances obtenues par chacun des critères d'évaluation sont affichées en rouge. On peut également noter que les méthodes basées sur le modèle de Markov homogène n'exploitent pas de variables d'entrée.

On note dans le tableau 3.9 que la variable de saisonnalité a été considérée dans les expérimentations effectuées sur les habitudes journalières, car celles-ci sont sujets à une périodicité hebdomadaire. Concernant les méthodes basées sur le modèle de Markov homogène et la régression logistique (MM, MixMM, LR et MixLR), on note une erreur de prévision élevée. On peut remarquer une meilleure performance de la méthode MixLR par rapport aux méthodes MM et MixMM. Au contraire, les méthodes basées sur le modèle de Markov non-homogène (JNMM, K-means+JNMM et MixJNMM) ont obtenu leurs meilleurs résultats en présence de l'ensemble des variables d'entrée (consommation, température, précipitation, calendrier et saisonnalité). Concernant les méthodes basées sur le modèle de Markov non-homogène, on note une précision plus précise en présence de la variable saisonnalité; cette variable s'est avérée nécessaire pour la prévision des habitudes journalières. La meilleure performance est obtenue en utilisant la méthode proposée (MixJNMM). Cela est due à la capacité de ce modèle à modéliser la dynamique de l'évolution des habitudes de consommation en tenant compte de l'influence des variables exogènes.

Concernant les habitudes hebdomadaires de consommation (voir tableau 3.10), étant donné l'horizon temporel des données, la variable saisonnalité n'a pas été considérée. Dans ce cas, nous n'observons plus de saisonnalités hebdomadaires ou mensuelles mais plutôt une saisonnalité annuelle des habitudes de consommation. Les résultats obtenus dans le cas d'états hebdomadaires sont assez proches de ceux obtenus dans le cas d'états journaliers. La méthode proposée fournit les meilleures performances en utilisant l'ensemble des variables de contexte.

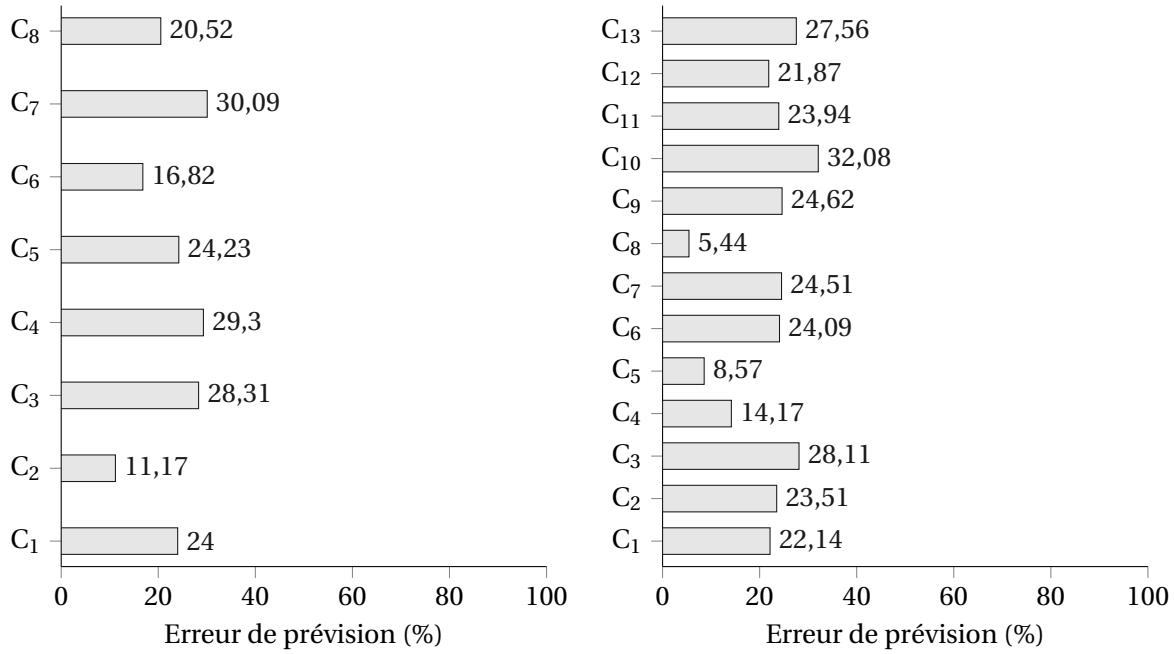
La figure 3.28 examine plus en détail le résultat de prévision obtenu en utilisant la méthode proposée (dans le cas des états journaliers et hebdomadaires). Dans cette figure, les performances en termes du taux d'erreur de prévision sont évaluées pour chaque classe identifiée. On peut noter que le taux d'erreur de prévision change de manière significative d'une classe à l'autre. Concernant les classes présentant des habitudes de consommation homogènes dans le temps (classe 2 dans

TABLEAU 3.9 – Tableau de comparaison des modèles dans le cas des profils journaliers durant 182 jours (base de test), MM : modèle de Markov homogène, MixMM : mélange de modèles de Markov homogènes, LR : modèle de régression logistique, MixLR : mélange de modèles de régressions logistiques, JNMM : modèle de Markov non-homogène, k-means+JNMM : modèle de Markov non-homogène au sein des classes identifiées par l'algorithme k-means, MixJNMM : mélange de modèles de Markov non-homogènes, Y : consommation, T : température, P : précipitation, C : évènements calendaires, S : saisonnalité

Models	**Inputs (e_{it})**	**ARI**	**Accuracy**	**Rappel**	**Précision**	**F-mesure**
MM		0.52	0.60	0.62	0.60	0.61
MixMM		0.55	0.63	0.63	0.64	0.63
LR	(T_{t-1},P_{t-1})	0.25	0.30	0.32	0.29	0.31
	(T_{t-1},P_{t-1},S_t)	0.29	0.34	0.34	0.36	0.35
	(Y_{t-1})	0.50	0.61	0.63	0.64	0.63
	(Y_{t-1},S_t)	0.50	0.61	0.63	0.58	0.60
	$(Y_{t-1},T_{t-1},P_{t-1},S_t)$	0.50	0.61	0.63	0.58	0.60
	(Y_{t-1},C_t,S_t)	0.52	0.61	0.64	0.61	0.62
	$(Y_{t-1},C_t,T_{t-1},P_{t-1})$	0.55	0.61	0.66	0.64	0.65
	$(Y_{t-1},C_t,T_{t-1},P_{t-1},S_t)$	0.55	0.61	0.67	0.65	0.66
MixLR	(T_{t-1},P_{t-1})	0.38	0.46	0.48	0.45	0.46
	(T_{t-1},P_{t-1},S_t)	0.40	0.48	0.50	0.49	0.49
	(Y_{t-1})	0.58	0.63	0.65	0.67	0.66
	(Y_{t-1},S_t)	0.59	0.65	0.66	0.67	0.66
	$(Y_{t-1},T_{t-1},P_{t-1},S_t)$	0.60	0.66	0.68	0.65	0.67
	(Y_{t-1},C_t,S_t)	0.57	0.62	0.64	0.63	0.63
	$(Y_{t-1},C_t,T_{t-1},P_{t-1})$	0.57	0.67	0.68	0.69	0.68
	$(Y_{t-1},C_t,T_{t-1},P_{t-1},S_t)$	0.63	0.67	0.68	0.66	0.67
JNMM	(T_{t-1},P_{t-1})	0.50	0.60	0.63	0.61	0.62
	(T_{t-1},P_{t-1},S_t)	0.65	0.73	0.72	0.70	0.71
	(Y_{t-1})	0.55	0.63	0.65	0.63	0.64
	(Y_{t-1},S_t)	0.66	0.73	0.72	0.73	0.72
	$(Y_{t-1},T_{t-1},P_{t-1},S_t)$	0.67	0.73	0.72	0.73	0.73
	(Y_{t-1},C_t,S_t)	0.68	0.74	0.72	0.72	0.72
	$(Y_{t-1},C_t,T_{t-1},P_{t-1})$	0.59	0.65	0.69	0.64	0.66
	$(Y_{t-1},C_t,T_{t-1},P_{t-1},S_t)$	0.70	0.74	0.73	0.74	0.73
K-means + JNMM	(T_{t-1},P_{t-1})	0.32	0.36	0.42	0.37	0.39
	(T_{t-1},P_{t-1},S_t)	0.48	0.52	0.56	0.52	0.54
	(Y_{t-1})	0.42	0.47	0.49	0.48	0.48
	(Y_{t-1},S_t)	0.62	0.66	0.68	0.64	0.66
	$(Y_{t-1},T_{t-1},P_{t-1},S_t)$	0.65	0.68	0.71	0.69	0.70
	(Y_{t-1},C_t,S_t)	0.67	0.69	0.71	0.72	0.71
	$(Y_{t-1},C_t,T_{t-1},P_{t-1})$	0.49	0.52	0.54	0.55	0.55
	$(Y_{t-1},C_t,T_{t-1},P_{t-1},S_t)$	0.68	0.72	0.74	0.71	0.73
MixJNMM	(T_{t-1},P_{t-1})	0.58	0.64	0.65	0.67	0.66
	(T_{t-1},P_{t-1},S_t)	0.65	0.74	0.72	0.74	0.73
	(Y_{t-1})	0.59	0.64	0.63	0.65	0.64
	(Y_{t-1},S_t)	0.68	0.75	0.74	0.75	0.75
	$(Y_{t-1},T_{t-1},P_{t-1},S_t)$	0.70	0.75	0.75	0.73	0.74
	(Y_{t-1},C_t,S_t)	0.71	0.76	0.74	0.72	0.73
	$(Y_{t-1},C_t,T_{t-1},P_{t-1})$	0.58	0.64	0.65	0.66	0.65
	$(Y_{t-1},C_t,T_{t-1},P_{t-1},S_t)$	0.75	0.78	0.79	0.81	0.80

TABLEAU 3.10 – Tableau de comparaison des modèles dans le cas des profils hebdomadaires durant 26 semaines (base de test), MM : modèle de Markov homogène, MixMM : mélange de modèles de Markov homogènes, LR : modèle de régression logistique, MixLR : mélange de modèles de régressions logistiques, JNMM : modèle de Markov non-homogène, *k*-means+JNMM : modèle de Markov non-homogène au sein des classes identifiées par l'algorithme de k-means, MixJNMM : mélange de modèles de Markov non-homogènes, Y : consommation, T : température, P : précipitation, C : évènements calendaires

Models	**Inputs (e_{it})**	**ARI**	**Accuracy**	**Rappel**	**Précision**	**F-mesure**
MM		0.36	0.65	0.67	0.67	0.67
MixMM		0.37	0.65	0.68	0.68	0.68
LR	(T_{t-1}, P_{t-1})	0.28	0.39	0.40	0.38	0.39
	(Y_{t-1})	0.60	0.67	0.68	0.66	0.67
	$(Y_{t-1}, T_{t-1}, P_{t-1})$	0.60	0.67	0.69	0.65	0.67
	(Y_{t-1}, C_t)	0.60	0.67	0.66	0.68	0.67
	$(Y_{t-1}, C_t, T_{t-1}, P_{t-1})$	0.60	0.67	0.68	0.66	0.67
MixLR	(T_{t-1}, P_{t-1})	0.44	0.56	0.58	0.55	0.56
	(Y_{t-1})	0.61	0.69	0.70	0.68	0.69
	$(Y_{t-1}, T_{t-1}, P_{t-1})$	0.61	0.69	0.69	0.67	0.68
	(Y_{t-1}, C_t)	0.61	0.69	0.68	0.68	0.68
	$(Y_{t-1}, C_t, T_{t-1}, P_{t-1})$	0.61	0.69	0.69	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_t)	0.61	0.73	0.72	0.72	0.72
	$(Y_{t-1}, C_t, T_{t-1}, P_{t-1})$	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_t)	0.62	0.74	0.75	0.74	0.74
	$(Y_{t-1}, C_t, T_{t-1}, P_{t-1})$	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_t)	0.64	0.79	0.78	0.76	0.77
	$(Y_{t-1}, C_t, T_{t-1}, P_{t-1})$	0.70	0.82	0.81	0.80	0.80



(a) Cas de séquences d'états journaliers

(b) Cas de séquences d'états hebdomadaires

FIGURE 3.28 – Erreur de prévision par classe en utilisant la méthode proposée MixJNMM et les variables d'entrées suivantes : température, précipitation et évènement calendaire

le cas d'états journaliers et classes 5 et 8 dans le cas d'états hebdomadaires), on remarque une erreur de prévision très faible. En revanche, cette erreur est plus importante dans le cas des classes présentant des habitudes de consommation plus variées.

3.4 Conclusion

Dans ce chapitre, nous avons abordé le problème lié au regroupement des séquences catégorielles qui présente un comportement dynamique dans le temps. Pour résoudre ce problème, nous nous sommes appuyés sur les modèles de mélange de lois. La spécificité du modèle proposé réside dans le fait que chaque composante du mélange est un modèle de Markov non homogène. Ce dernier permet de modéliser l'évolution dynamique du comportement tout en tenant compte de l'influence des facteurs exogènes. L'approche proposée a fait l'objet d'une évaluation sur des données synthétiques qui a mis en évidence sa capacité à modéliser les différents scénarios conçus.

En appliquant cette méthode sur des séquences d'habitudes de consommation réelles, et en se basant sur le critère BIC associé, 8 classes de compteurs dans le cas des habitudes journalières et 13 classes de compteurs dans le cas des habitudes hebdomadaires ont été obtenues. L'examen des classes a donné lieu à des interprétations réalistes, et a permis d'associer chaque classe à une catégorie socio-professionnelle. Une fois la classification effectuée, les paramètres estimés au sein de chaque classe ont été exploités pour effectuer des prévisions de futures habitudes de consommation. Les résultats obtenus ont montré la pertinence de la méthode proposée pour la prévision des habitudes de consommation.

Afin d'analyser les changements dans l'évolution des habitudes de consommation au sein des classes obtenues, le chapitre suivant s'intéresse au problème de la détection de changements communs à un ensemble de compteur. Ainsi, les groupes obtenus dans ce chapitre seront utilisés comme entrée pour les algorithmes développés dans le chapitre 4.

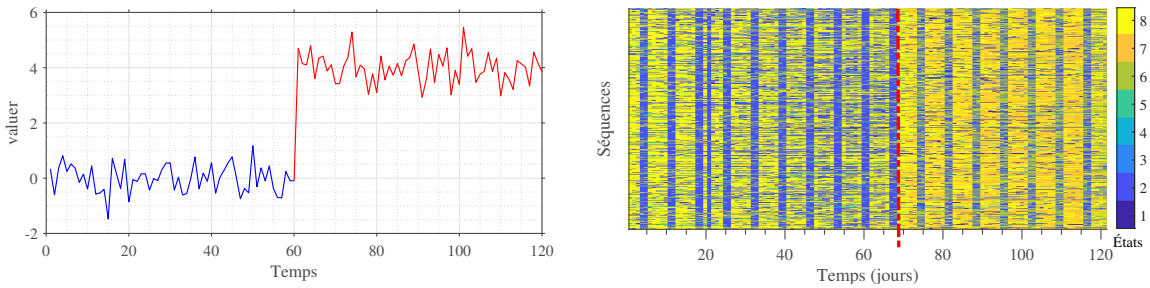
Chapitre 4

Détection de changement dans un panel de séquences catégorielles

Sommaire

4.1 Introduction

La détection de changement dans un processus stochastique est un problème important et répandu dans divers domaines. L'objectif global des méthodes de détection est d'estimer les instants de changement dans la distribution des données observées. Plusieurs articles et ouvrages ont abordé ce sujet (BASSEVILLE, 1988; BASSEVILLE et collab., 1993; GUSTAFSSON, 2000; ISERMANN, 1984). Les problèmes de détection peuvent porter sur des séquences d'observations numériques monovariées (voir figure 4.1a), ou plus généralement sur des séquences multivariées à caractère discret ou continu. Ils peuvent également porter sur des objets statistiques plus complexes tels que des courbes, des images ou des graphes. Dans cette thèse, on s'intéresse au problème de la détection de changements communs à un ensemble de séquences catégorielles (voir figure 4.1b) et ce, de manière séquentielle.



(a) Changement détecté dans une série continue

(b) Changement détecté dans de multiples séquences

FIGURE 4.1 – Exemples de la détection de changement. Détection d'un changement du comportement dans une série continue (a), détection de changement dans de multiples séquences catégorielles (b). Les changements détectés sont signalés en rouge.

Rappelons que dans le chapitre précédent, la méthodologie proposée avait permis de regrouper les séquences catégorielles (compteurs) en classes; chaque classe étant constituée de compteurs caractérisés par une dynamique propre d'habitudes de consommation. Dans le contexte des réseaux d'eau, l'évolution du comportement des usagers peut être liée à plusieurs facteurs tels que des changements démographiques, des changements de saison, des changements de comportement intrinsèques ou également des fuites d'eau. L'analyse de cette évolution et l'identification des changements de comportement pourrait permettre aux compagnies d'eau de mieux surveiller le réseau et ainsi de mieux adapter leur service en fonction de la demande.

Dans ce cadre, ce chapitre présente une méthode de détection de changements communs à un ensemble de compteurs. En d'autres mots, on s'intéresse, ici, au partitionnement des habitudes de consommation en segments temporels contigus. Pour ce faire, nous proposons une méthode basée sur les tests séquentiels du rapport de vraisemblance, qui sont fondés eux-mêmes sur les modèles de Markov non-homogènes. Ce choix du modèle nous a semblé pertinent, car il permet de modéliser la régularité temporelle des séquences en tenant compte de l'influence de variables exogènes. Pour pouvoir détecter les différents types de changement au sein d'un ensemble de séquences caractérisées par un comportement non-stationnaire au fil du temps, nous proposons d'utiliser un seuil adaptatif, qui est estimé en utilisant des simulations de Monte Carlo.

Dans la suite de cette section, un état de l'art des approches les plus utilisées dans le domaine de la détection de changement est d'abord fourni. Celles-ci visent, plus particulièrement, le problème de la détection de changement dans les séries temporelles numériques. On présente également quelques tests séquentiels de référence. Ensuite, nous présentons quelques méthodes permettant de détecter les changements dans les séquences catégorielles. Finalement, la méthode de détection de changement proposée est introduite et les résultats obtenus sont présentés.

4.1.1 Méthodes de détection de changement

Dans cette section, nous présentons dans un cadre général, les méthodes de détection de changement qui sont utilisées dans divers domaines statistiques. Les méthodes de détection basées sur les tests d'hypothèses sont d'abord introduites et une extension de celles-ci permettant de traiter les données de manière séquentielle est présentée. Ensuite, les approches basées sur la segmentation contiguë des séquences sont décrites. Cette section commence par une introduction des tests d'hypothèses.

Introduction aux tests d'hypothèses

Les tests d'hypothèses sont généralement utilisés pour examiner la véracité d'une affirmation concernant la distribution d'un échantillon de données observées. Dans le cadre de la détection de changement, cette affirmation (existence d'un point de changement) est vérifiée par deux hypothèses opposées concernant la distribution des données : l'hypothèse nulle et l'hypothèse alternative. En général, l'hypothèse nulle traduit l'absence de points de changement et l'hypothèse alternative stipule la présence d'un ou plusieurs points de changement. On distingue, en général, deux types de test qui sont : les *tests hors ligne* et les *tests en ligne*. Dans le cas d'un test hors ligne, on considère que la séquence de données est entièrement observée et dans le cas d'un test en ligne, on considère que les données parviennent au fil du temps. La détection hors ligne pour une séquence observée $y_1, \ldots, y_T$ et dans le cas où les données sont indépendantes et identiquement distribuées avant et après le changement s'appuie sur les hypothèses de test suivantes :

$$\begin{cases} \mathrm{H}_0 : & y_t \sim \mathrm{P}_{\tilde{\theta}_0} \quad \forall\, t = 1, \ldots, \mathrm{T} \\ \mathrm{H}_1 : & y_t \sim \mathrm{P}_{\theta_0} \quad \forall\, t = 1, \ldots, \tau - 1, \\ & y_t \sim \mathrm{P}_{\theta_1} \quad \forall\, t = \tau, \ldots, \mathrm{T} \end{cases} \tag{4.1}$$

où τ indique l'instant où la distribution des données pourrait présenter un changement et θ_0 et θ_1 sont les paramètres associés à cette distribution avant et après le changement.

Dans le cas de l'acquisition séquentielle des données, on peut s'appuyer sur la suite des problèmes suivants :

$$\forall\, t', \quad \begin{cases} \mathrm{H}_0 : & y_t \sim \mathrm{P}_{\tilde{\theta}_0} \quad \forall\, t = 1, \ldots, t' \\ \mathrm{H}_1 : & y_t \sim \mathrm{P}_{\theta_0} \quad \forall\, t = 1, \ldots, \tau - 1 \\ & y_t \sim \mathrm{P}_{\theta_1} \quad \forall\, t = \tau, \ldots, t'. \end{cases} \tag{4.2}$$

L'objectif visé par les test séquentiels est de réduire le délai de détection tout en garantissant un nombre faible de fausses alarmes. Le délai de détection est défini comme la durée entre la détection à l'instant t et le changement τ.

La section suivante présente quelques tests séquentiels qui permettent de détecter de multiples points de changement.

Méthode de détection de changement en ligne basée sur le rapport de vraisemblance

Les méthodes de détection de changement en ligne de séries temporelles visent à estimer les points de changement au fur et à mesure que les données parviennent. On note $(y_1, \ldots, y_t, \ldots, y_T)$, une série temporelle observée de manière séquentielle dans le temps. Les méthodes en ligne supposent généralement que les paramètres avant le point de changement θ_0 sont connus. Elles sont

basées de manière générale, sur le logarithme du rapport de vraisemblance (LR), qui s'écrit :

$$\begin{aligned} S_1^T(\tau) &= \log \frac{P(y_1,\ldots,y_T|\tau,\theta_0,\theta_1)}{P(y_1,\ldots,y_T|\theta_0)} \\ &= \log \frac{P(y_1,\ldots,y_{\tau-1}|\theta_0)\cdot P(y_\tau,\ldots,y_T|\theta_1)}{P(y_1,\ldots,y_{\tau-1}|\theta_0)\cdot P(y_\tau,\ldots,y_T|\theta_0)} \\ &= \log \frac{P(y_\tau,\ldots,y_T|\theta_1)}{P(y_\tau,\ldots,y_T|\theta_0)}. \end{aligned} \tag{4.3}$$

S est également appelé la *statistique de test*. Cette statistique est calculée pour $1 \leq \tau \leq T$. Pour détecter les points de changement, on s'appuie sur la règle de décision optimale d qui est la suivante :

$$d = \begin{cases} 0 & si \quad S_1^T(\tau) < h; \quad H_0 \text{ est sélectionné} \\ 1 & si \quad S_1^T(\tau) \geq h; \quad H_1 \text{ est sélectionné,} \end{cases} \tag{4.4}$$

où h est le seuil déterminé de manière empirique à partir de données observées. Ainsi, l'instant de changement est donnée par :

$$\hat{\tau} = \arg\max_{\tau} S_1^T(\tau). \tag{4.5}$$

Dans le cas où la distribution des données avant et après le changement est connue, le test de CUSUM proposé dans PAGE (1954) peut être utilisé. Ce test est décrit dans la section suivante.

Test du CUSUM

Le test de la somme cumulée (CUSUM) suppose que les paramètres des lois avant et après changement (θ_0 et θ_1) sont connus, hypothèse sous laquelle on peut établir que ce test minimise le délai moyen de détection pour un taux de fausses alarmes fixé (LORDEN et collab., 1971; MOUSTAKIDES et collab., 1986). Le logarithme du rapport de vraisemblance pour ce test est donné par :

$$s_1^t = \log \frac{P(y_1,\ldots,y_t|\tau,\theta_0,\theta_1)}{P(y_1,\ldots,y_t|\theta_0)}, \tag{4.6}$$

et sa somme cumulée jusqu'à l'instant t est définie par :

$$S_1^t = \sum_{t'=1}^{t} s_1^{t'}. \tag{4.7}$$

Ce test consiste à calculer de manière récursive la statistique du rapport de vraisemblance entre l'hypothèse alternative de présence d'un point de changement et l'hypothèse nulle d'absence de point de changement :

$$S_1^t = S_1^{t-1} + s_1^t. \tag{4.8}$$

Comme indiqué par BASSEVILLE et collab. (1993), la statistique de test est comparée à une valeur de seuil h positive. Donc on peut réécrire (4.8) comme suit :

$$g_1^t = \{g_1^{t-1} + s_1^t\}^+, \tag{4.9}$$

où $\{\cdot\}^+ = sup(\cdot, 0)$. Si $g_t > h$ un changement est signalé et l'instant correspondant est estimé par :

$$\hat{\tau} = \arg\min_{1\leq\tau\leq t'} S_1^t(\tau - 1). \tag{4.10}$$

Dans le cas où l'on cherche à détecter plusieurs changements, il suffit de réinitialiser ce détecteur dès lors qu'un point de changement a été décelé. Dans les situations réelles, où les paramètres des modèles avant et après changement sont inconnus, on peut s'appuyer sur la statistique du rapport de vraisemblance généralisé dont le calcul incorpore l'estimation par la méthode du maximum de vraisemblance des paramètres inconnus (BASSEVILLE et collab., 1993).

Test du GLR

Dans les situations réelles, la distribution des données après un point de changement $P(\mathbf{y};\theta_1)$ est inconnue. Pour pouvoir mener le test dans le cas où θ_1 est inconnu, le test du rapport de vraisemblance généralisé (*Generalized likelihood ratio*) proposé par LORDEN et collab. (1971) peut être utilisé. Ce test considère que le paramètre θ_0 est connu. Comme dans le cas de test du CUSUM, le rapport du vraisemblance s'écrit :

$$S_1^t(\tau,\theta_1) = \log\frac{P(y_1,\ldots,y_t|\tau,\theta_0,\theta_1)}{P(y_1,\ldots,y_t|\theta_0)} \tag{4.11}$$

Dans ce cas, pour estimer les paramètres inconnus qui sont l'instant du changement et la valeur du paramètre après le changement, on s'appuie sur la méthode du maximum de vraisemblance définie dans cette situation par la double maximisation suivante :

$$\begin{aligned} g_1^t &= \max_{1\leq\tau\leq t}\sup_{\theta_1} S_1^t(\tau,\theta_1) \\ &= \max_{1\leq\tau\leq t}\sum_{i=\tau}^{k}\log\frac{P(y_i|\hat{\theta}_1)}{P(y_i,\theta_0)}, \end{aligned} \tag{4.12}$$

où θ_0 est supposé connu et $\hat{\theta}_1$ correspond à l'estimation par la maximisation de la vraisemblance $P(y_\tau,\ldots,y_t|\theta)$.

Segmentation en utilisant le critère BIC

Le problème de la détection de changement dans les séries temporelles peut être considéré comme un problème du partitionnement des séries en segments contigus dans le temps. Dans le cadre général, deux modèles sont estimés à partir des données observées : un modèle avec un vecteur de paramètres et un autre modèle avec deux vecteurs de paramètre différents. Finalement, le critère BIC (SCHWARZ et collab., 1978b) permet de sélectionner le modèle adapté.

Une méthode générique a été proposée par CHEN et collab. (1998) pour la segmentation séquentielle de flux audio, permettant ainsi de détecter les changements de l'orateur et ceux liés à un changement dans l'environnement. Dans cette étude, on considère que les données $(y_1,\ldots,y_T)$ sont distribuées suivant une densité gaussienne et un changement est détecté lorsque les paramètres de la distribution changent. Les tests d'hypothèses sont utilisés pour vérifier l'occurrence d'un changement à l'instant τ :

$$\begin{cases} H_0 : y_1,\ldots,y_T \sim \mathcal{N}(\mu,\Sigma) \\ H_1 : y_1,\ldots,y_\tau-1 \sim \mathcal{N}(\mu_1,\Sigma_1); \\ \quad\;\; y_\tau,\ldots,y_T \sim \mathcal{N}(\mu_2,\Sigma_2) \end{cases} \tag{4.13}$$

La différence entre les critères BIC issus de ces deux modèles s'écrit :

$$BIC(\tau) = N\log|\Sigma| - N_1\log|\Sigma_1| - N_2\log|\Sigma_2| - \lambda\vartheta, \tag{4.14}$$

où Σ, Σ_1 et Σ_2 sont des matrices de covariance estimées respectivement sur l'ensemble des données, sur le segment avant le changement $\{y_1,\ldots,y_{\tau-1}\}$ et sur le segment après le changement $\{y_\tau,\ldots,y_T\}$. La taille des segments est indiquée par N, N_1 et N_2, ϑ désigne le nombre de paramètres (pénalisation) et λ est le poids associé. Ce poids peut être vu comme un seuil intégré qui doit être ajusté. Si la valeur de BIC est positive pour un instant τ, le modèle avec deux gaussiennes est favorisé. Donc, on détecte un point de changement si

$$\{\max_{\tau} BIC(\tau)\} > 0. \tag{4.15}$$

L'estimation du maximum de vraisemblance pour le point de changement s'écrit :

$$\hat{\tau} = \arg\max_{\tau} \text{BIC}(\tau). \tag{4.16}$$

Dans le cadre de détection de multiples points de changement, une fenêtre de taille croissante est utilisée. Lorsqu'un point de changement est détecté, cette fenêtre est réinitialisée après ce dernier ($\hat{\tau}+1$).

Dans le domaine ferroviaire, SAMÉ et GOVAERT (2012) ont proposé également une méthodologie similaire pour la détection de changement dans les séries multivariées. La particularité de cette méthode est qu'il n'est pas nécessaire de fixer une valeur de seuil. Les modèles de mélange sont exploités pour définir les hypothèses de test. Selon cette approche, les données sont distribuées suivant un modèle avec une composante régressive sous l'hypothèse nulle, et suivant un mélange de deux composantes régressives sous l'hypothèse alternative. Pour pouvoir adapter les modèles de mélange au partitionnement temporel des séries, les composantes de celui-ci sont des fonctions du temps :

$$f(\mathbf{y}_t|\boldsymbol{\theta}) = \sum_{k=1}^{\text{K}} p_k(t;\boldsymbol{\alpha})\mathcal{N}_d(\mathbf{y}_t|\boldsymbol{\mu}_k(t),\boldsymbol{\Sigma}_k). \tag{4.17}$$

Les proportions du mélange sont des transformations logistiques des fonctions linéaires du temps, ce qui permet de garantir un partitionnement des séquences en segments contigus.

Finalement, la différence entre les critères BIC issus de chaque modèle permet d'identifier les points de changement.

Une version séquentielle de ce modèle est également proposée, qui utilise une approche similaire à celle proposée par CHEN et collab. (1998). Suivant cette approche, après chaque détection la fenêtre croissante est réinitialisée à l'instant $a = \min\{a < t < b; p_1(t;\boldsymbol{\alpha}) < 0.5\}$.

Autres méthodes

Pour détecter en ligne des anomalies dans un flux de données fonctionnelles issu du monitoring d'une flotte d'autobus, CHEIFETZ et collab. (2013) ont initié une méthode basée sur un test du rapport de vraisemblance généralisé construit à partir du modèle RHLP. Cette approche a permis de détecter des changements de comportement globaux et a fourni des détecteurs locaux pour caractériser des anomalies.

Une méthode en ligne basée sur l'analyse de l'erreur de prédiction a été proposée dans AHMAD et collab. (2017) pour la détection en temps réel de changement dans des flux de données. Un critère d'erreur de type Similarité Cosinus, calculé entre les données et la prédiction issue de la méthode « Hierarchical Temporal Memory » (HTM), permet de détecter les anomalies. Pour pouvoir détecter plus efficacement des changements en présence de données bruitées, les auteurs proposent de modéliser la distribution des erreurs de prédiction et utiliser celle-ci pour vérifier la cohérence de l'anomalie à chaque instant. Il est supposé que cette distribution est normale et ses paramètres sont mis à jour en utilisant des fenêtres sur l'historique des erreurs de prédiction. Cette méthode s'est révélée pertinente pour la détection spatiale et temporelle des changements dans les situations de données fortement bruitées.

Dans le cadre de l'estimation bayésienne de points de changement, une méthode de segmentation en ligne basée sur le calcul récursif exact de la distribution a posteriori des points de changement a été proposée dans FEARNHEAD et LIU (2007). Une extension de ce modèle, incluant l'estimation par la méthode du maximum de vraisemblance des paramètres, a été développée dans CARON et collab. (2012).

4.1.2 Méthodes de détection de changement dans une séquence catégorielle

Dans le cas des séries catégorielles (ordinale, binomiale, multinomiale), une méthode en ligne basée sur les tests CUSUM est proposée par HÖHLE (2010). Dans le cas des séries binomiales et bêta-binomiale, on suppose que celles-ci sont distribuées suivant un modèle de régression logistique. Cette approche permet de détecter, de manière en ligne, un changement structurel dans l'intercepte du modèle estimé à l'aide de l'approche de la somme cumulée. Une extension de ce modèle en utilisant un modèle de régression logistique mulinomiale a également été introduite.

Une extension des travaux menés par FOKIANOS et collab. (2014) pour la détection de changement dans les séries binaires est proposée par GOMBAY et collab. (2017) qui traite les séries temporelles multinomiales. La méthode proposée est basée sur les tests d'hypothèses et on suppose que les données avant et après le changement sont distribuées suivant un modèle de régression logistique multinomiale (MLR). La statistique de test est basée sur le dérivé de la fonction de log-vraisemblance. La comparaison de ce score avec un seuil fournit les points de changement.

Dans le contexte des données catégorielles multivariées, une étude similaire menée par LI et collab. (2013), propose une méthode pour la détection de changement directionnelle qui est basée sur les modèles linéaires logarithmiques.

Détection de changement à l'aide des modèles de Markov caché

Les modèles de Markov sont le plus souvent utilisés pour la modélisation des séquences temporelles. Ils permettent de modéliser la relation de dépendance entre les observations au fil du temps. Dans une étude menée par LUONG et collab. (2012), les modèles de Markov caché (HMM) sont utilisés pour une segmentation temporelle des séquences. Dans ce contexte, chaque transition entre segments correspond à un point de changement, et les observations dans chaque segment sont modélisées par un modèle de Markov homogène. La segmentation d'une séquence $(z_1,\dots,z_T)$ en K partitions est définie par :

$$P(z_1,\dots,z_T|S_1,\dots,S_T;\theta) = \prod_{t=1}^{T} P(z_t|S_t;\theta) = \prod_{k=1}^{K} \prod_{t,S_t=k} P(y_t|S_t = k;\theta_k), \tag{4.18}$$

où $(S_1,\dots,S_T)$ désigne l'ensemble des états cachés associés avec S_t étant l'indexe d'un segment à l'instant t, $P(z_t|S_t = k)$ est la distribution des données observées dans le segment k qui est un modèle de Markov homogène et $\theta = (\theta_1,\dots,\theta_K)$ est le vecteur des paramètres du modèle. La densité jointe des observations et des segments est donnée par :

$$P(z_1,\dots,z_T,S_1,\dots,S_T) = P(S_1)P(z_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(z_t|S_t). \tag{4.19}$$

L'algorithme *forward-backward* (RABINER, 1989) permet d'estimer les paramètres du modèle et d'identifier l'ensemble des segments.

4.1.3 Problème de la détection de changements dans un ensemble de séquences catégorielles

La figure 4.2 montre un exemple illustratif de la détection de changement dans le cas des séries catégorielles. Chaque ligne correspond à une séquence d'habitudes de consommation (compteur) observée dans le temps et chaque colonne correspond aux habitudes hebdomadaires de consommation adoptées par l'ensemble des séquences. Les habitudes de consommation appelées également *états* sont affichées en utilisant leur labels (8 modalités). On peut noter deux points de

changement communs à l'ensemble des séquences (indiqués par les pointillés rouges). Dans la suite, nous proposons une méthodologie permettant de détecter les changements communs à un ensemble de séquences catégorielles.

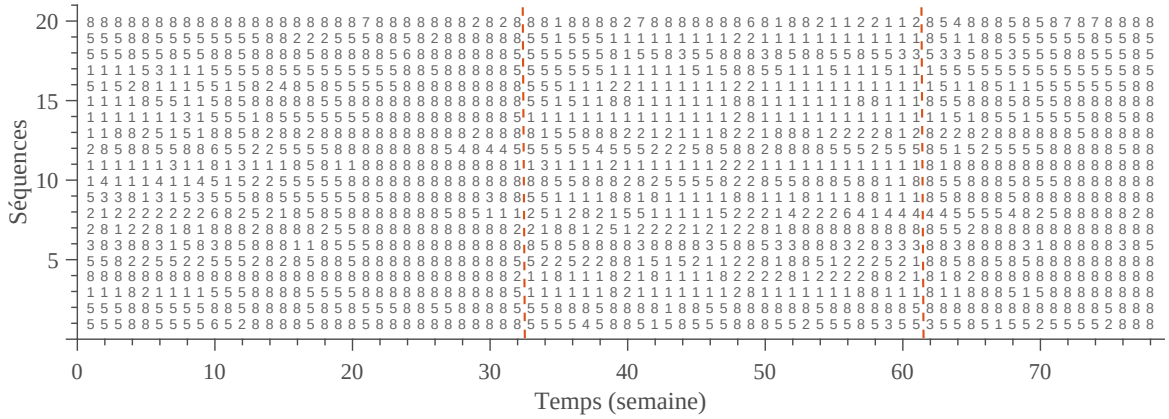


FIGURE 4.2 – Exemple de changement de changements détectés dans les habitudes hebdomadaires de consommation d'un ensemble de 20 compteurs; les chiffres indiquent les habitudes codées et les traits rouges en pointillé correspondent aux instants de changement d'habitudes

4.2 Méthodologie proposée pour la détection de changements dans un ensemble de séquences

La méthode proposée s'appuie sur les tests séquentiels d'hypothèses JUNG et collab. (2004) pour détecter un ou plusieurs points de changement communs à un ensemble de séquences catégorielles. Cela implique que ces dernières aient une évolution similaire dans le temps. Pour rester conforme à cette hypothèse, nous nous sommes focalisés sur les séquences issues des classes identifiées dans le chapitre précédent (voir figures 3.17b et 3.18b). Plus précisément, l'approche proposée est basée sur le test du rapport de vraisemblance généralisé. Les hypothèses de test sont les suivantes :

$$\begin{cases} \mathrm{H}_0 : & \mathcal{M} = \tilde{\mathcal{M}}_0 \quad \text{pour} \quad 1 \leq t \leq \mathrm{T} \\ \mathrm{H}_\mathrm{A} : & \mathcal{M} = \mathcal{M}_0 \quad \text{pour} \quad 1 \leq t < \tau \\ & \mathcal{M} = \mathcal{M}_1 \quad \text{pour} \quad \tau \leq t \leq \mathrm{T} \end{cases} \tag{4.20}$$

où τ indique un instant de changement, $\mathcal{M}_0$ et $\mathcal{M}_1$ sont respectivement des modèles associés aux segments avant et après le point de changement et $\tilde{\mathcal{M}}_0$ est le modèle associé à la séquence entière. Concernant le choix du modèle, dans cette thèse nous avons opté pour les modèles de Markov non homogènes. Ce choix a été motivé par la capacité des modèles de Markov à modéliser le comportement dynamique des séquences temporelles. Il permet également de tenir compte des facteurs exogènes, ce qui permettra dans la suite de caractériser les changements détectes.

L'hypothèse H_0 (voir (4.20)) est l'hypothèse sous laquelle les données sont distribuées suivant le même modèle de Markov de paramètres $\tilde{\boldsymbol{\theta}}_0 = (\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\beta}}_0)$ sur toute la période (absence de changement). L'hypothèse alternative H_A considère que la distribution des données pourrait faire l'objet à un changement à partir d'un instant de temps noté τ. Ainsi, les données sont distribuées suivant deux modèles de Markov non homogènes de paramètres $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ et $\boldsymbol{\theta}_1 = (\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)$, avant et après le point de changement τ. La figure 4.3 montre les intervalles de temps associés aux hypothèses de test, en supposant que la séquence est observée jusqu'à l'instant T. Dans la suite de cette section, on désignera par $\mathbf{z} = (z_{it})_{1 \leq i \leq n, 1 \leq t \leq \mathrm{T}}$ un ensemble de n séquences de longueur T (nombre de jours ou nombre de semaines) et par $\mathbf{u} = (u_{it})_{1 \leq i \leq n, 1 \leq t \leq \mathrm{T}}$ les variables de contexte associées.

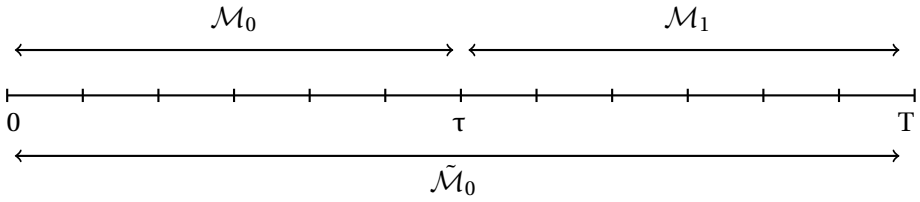


FIGURE 4.3 – Représentation des intervalles de temps associés aux hypothèses de test

Pour décider entre les deux hypothèses décrites ci-dessus, on s'appuie sur la statistique du rapport de vraisemblance généralisé, qui s'écrit de manière générale,

$$\Lambda_1^{\mathrm{T}}(\tau,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\tilde{\boldsymbol{\theta}}_0) = \log\left[\frac{\mathrm{P}_{\mathcal{M}_0}(z_1,\ldots,z_{\tau-1}|\mathbf{u}_1,\ldots,\mathbf{u}_{\tau-1})\mathrm{P}_{\mathcal{M}_1}(z_\tau,\ldots,z_{\mathrm{T}}|\mathbf{u}_\tau,\ldots,\mathbf{u}_{\mathrm{T}})}{\mathrm{P}_{\tilde{\mathcal{M}}_0}(z_1,\ldots,z_{\mathrm{T}}|\mathbf{u}_1,\ldots,\mathbf{u}_{\mathrm{T}})}\right], \tag{4.21}$$

où $(z_1,\ldots,z_{\mathrm{T}})$ désigne une séquence d'états pour un compteur, $(\mathbf{u}_1,\ldots,\mathbf{u}_{\mathrm{T}})$ indique les variables de contexte associées, $\mathrm{P}_{\mathcal{M}_0}$ et $\mathrm{P}_{\mathcal{M}_1}$ désignent les distributions de probabilité conditionnelle des segments situés avant et après le point de changement τ et $\mathrm{P}_{\tilde{\mathcal{M}}_0}$ est la distribution de probabilité conditionnelle portant sur la séquence entière. Cette statistique est calculée pour différentes valeurs de τ ($1 \le \tau \le \mathrm{T}$). En développant l'équation (4.21), on obtient :

$$\begin{aligned}\Lambda_1^{\mathrm{T}}(\tau,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\tilde{\boldsymbol{\theta}}_0) &= \sum_{i=1}^{n}\log\mathrm{P}_{\mathrm{M}_0}(z_{i1}|\mathbf{u}_{i1}) + \sum_{i=1}^{n}\sum_{t=2}^{\tau-1}\log\mathrm{P}_{\mathrm{M}_0}(z_{it}|z_{it-1},\mathbf{u}_{it}) \\ &+ \sum_{i=1}^{n}\log\mathrm{P}_{\mathrm{M}_1}(z_{i\tau}|\mathbf{u}_{i\tau}) + \sum_{i=1}^{n}\sum_{t=\tau+1}^{\mathrm{T}}\log\mathrm{P}_{\mathrm{M}_1}(z_{it}|z_{it-1},\mathbf{u}_{it}) \\ &- \sum_{i=1}^{n}\log\mathrm{P}_{\tilde{\mathrm{M}}_0}(z_{i1}|\mathbf{u}_{i1}) - \sum_{i=1}^{n}\sum_{t=2}^{\mathrm{T}}\log\mathrm{P}_{\tilde{\mathrm{M}}_0}(z_{it}|z_{it-1},\mathbf{u}_{it}).\end{aligned} \tag{4.22}$$

En remplaçant les distributions de probabilité par les modèles de Markov non homogènes correspondants, on peut écrire :

$$\begin{aligned}\Lambda_1^{\mathrm{T}}(\tau,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\tilde{\boldsymbol{\theta}}_0) &= \sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} z_{i1k}\log\pi_k(\mathbf{u}_{i1};\boldsymbol{\alpha}_0) + \sum_{i=1}^{n}\sum_{t=2}^{\tau-1}\sum_{k,\ell=1}^{\mathrm{K}} z_{itk}z_{i(t-1)\ell}\pi_{k\ell}(\mathbf{u}_{it};\boldsymbol{\beta}_{0\ell}) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} z_{i\tau k}\log\pi_k(\mathbf{u}_{i\tau};\boldsymbol{\alpha}_1) + \sum_{i=1}^{n}\sum_{t=\tau+1}^{\mathrm{T}}\sum_{k,\ell=1}^{\mathrm{K}} z_{itk}z_{i(t-1)\ell}\log\pi_{k\ell}(\mathbf{u}_{it};\boldsymbol{\beta}_{1\ell}) \\ &- \left(\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} z_{i1k}\log\pi_k(\mathbf{u}_{i1};\tilde{\boldsymbol{\alpha}}_0) + \sum_{i=1}^{n}\sum_{t=2}^{\mathrm{T}}\sum_{k,\ell=1}^{\mathrm{K}} z_{itk}z_{i(t-1)\ell}\pi_{k\ell}(\mathbf{u}_{it};\tilde{\boldsymbol{\beta}}_{0\ell})\right),\end{aligned} \tag{4.23}$$

où les variables indicatrices z utilisées sont les suivantes :

- $z_{i1k} = 1$ si $z_{i1} = k$; sinon $z_{i1k} = 0$,
- $z_{itk}z_{i(t-1)\ell} = 1$ si $z_{it} = k$ et $z_{i(t-1)} = \ell$; sinon $z_{itk}z_{i(t-1)\ell} = 0$,

et $\pi_k(\cdot)$ est la probabilité initiale de paramètres $\boldsymbol{\alpha}$ et $\pi_{k\ell}(\cdot)$ désigne la probabilité de transition de l'état ℓ vers l'état k de paramètres $\boldsymbol{\beta}$. Celles-ci sont définies comme suite :

$$\pi_k(\mathbf{u}_{i,1};\boldsymbol{\alpha}) = \mathrm{P}(z_{i,1} = k \mid \mathbf{u}_{i1}) = \frac{e^{\boldsymbol{\alpha}_k^\top \mathbf{u}_{i,1}}}{\sum_{\ell=1}^{\mathrm{K}} e^{\boldsymbol{\alpha}_\ell^\top \mathbf{u}_{i,1}}}, \tag{4.24}$$

$$\pi_{k,\ell}(\mathbf{u}_{i,t};\boldsymbol{\beta}_\ell) = \mathrm{P}(z_{i,t} = k \mid z_{i,t-1} = \ell, \mathbf{u}_{it}) = \frac{e^{\boldsymbol{\beta}_{k,\ell}^\top \mathbf{u}_{i,t}}}{\sum_{\ell=1}^{\mathrm{K}} e^{\boldsymbol{\beta}_{k,\ell}^\top \mathbf{u}_{i,t}}}. \tag{4.25}$$

Le vecteur des variables d'entrées $\mathbf{u}_{it} = (1, \mathrm{Y}_t, \mathrm{T}_t, \mathrm{P}_t, \mathrm{E}_t, \mathrm{C}_1(t), \mathrm{S}_1(t), \ldots, \mathrm{C}_q(t), \mathrm{S}_q(t))$ est constitué du volume d'eau consommé (issu des données brutes non normalisées, Y_T) et de l'ensemble de variables exogènes telles que la température (exprimée en degrés centigrades, T_t), la précipitation (en millimètres, P_t) et le calendrier (une variable binaire indiquant les périodes de vacances scolaires et les jours fériés, E_t), qui peuvent avoir une influence sur les habitudes de consommation. Les variables $\mathrm{C}_q(t)$ et $\mathrm{S}_q(t)$ sont des termes trigonométriques (voir équation 3.35) qui permettent de tenir compte de la saisonnalité des habitudes de consommation.

Le vecteur $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_{0\ell}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_{1\ell}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\beta}}_{0\ell}, \tau)$ désigne l'ensemble des paramètres de la méthode de détection de changement proposée (paramètres des modèles de Markov et point de changement τ). Les paramètres sont estimés à l'aide de la méthode du maximum de vraisemblance :

$$\Lambda_\mathrm{T} = \max_{\tau, (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1), \tilde{\boldsymbol{\theta}}_0} \Lambda_1^\mathrm{T}(\tau). \tag{4.26}$$

Pour un point de changement τ, les paramètres de modèles de Markov $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ et $\tilde{\boldsymbol{\theta}}_0$ sont estimés respectivement dans les intervalles $[1;\tau-1]$, $[\tau;\mathrm{T}]$ et $[1;\mathrm{T}]$. Une fois que les paramètres sont estimés, les statistiques $\Lambda_1^\mathrm{T}(\tau)$ et Λ_T sont déduites. La procédure d'estimation des paramètres, qui est fondée sur l'algorithme IRLS (GREEN, 1984), est détaillée dans l'annexe A. Finalement, le point de changement peut être estimé en utilisant la formule suivante :

$$\hat{\tau} = \arg\max_{\tau} \left[\max_{(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1), \tilde{\boldsymbol{\theta}}_0} \Lambda_1^\mathrm{T}(\tau) \right]. \tag{4.27}$$

La règle de décision suivante permet de déterminer si le point de changement estimé correspond à un changement de comportement ou non :

$$d = \begin{cases} 0 & \text{si} \quad \Lambda_\mathrm{T} < h \\ 1 & \text{si} \quad \Lambda_\mathrm{T} \geq h. \end{cases} \tag{4.28}$$

Pour prendre une décision, la statistique de test (calculée en utilisant l'équation (4.26)) est comparée à un seuil h. Si la valeur de la statistique de test dépasse le seuil, on considère l'existence d'un point de changement $\hat{\tau}$ ($d = 1$). Ce dernier peut être obtenu en utilisant l'équation 4.27. Dans le cas contraire, la méthode proposée ne détecte aucun changement.

Le choix d'un seuil approprié semble être primordial et exige le plus souvent des connaissances expertes du phénomène analysé. Puisque le choix d'un seuil inadapté pourrait conduire à un taux élevé de fausses alarmes ou bien à la non-détection des points de changement (faux négatifs). Pour déterminer la valeur du seuil, nous optons pour une approche statistique basée sur les simulations de type Monte Carlo. La section suivante décrit la procédure d'estimation du seuil.

4.2.1 Choix du seuil de détection

Cette section présente deux approches différentes pour déterminer la valeur du seuil. La première consiste à fixer un seuil de manière empirique pour détecter un ensemble de points de changement, et la deuxième propose un seuil adaptatif qui varie en fonction des données observées. De manière générale, la valeur de seuil sélectionnée contribue à établir un compromis raisonnable entre le délai de détection et le taux de fausses alarmes.

4.2.1.1 Seuil fixe

Cette approche considère que le seuil est estimé une seule fois à partir d'une séquence de données initiales, et cette valeur reste invariable pour la détection de points de changement sur le reste

de la période. Ce type de seuil nécessite le plus souvent une intervention des experts du domaine ou une validation croisée hors ligne avant la procédure de détection pour confirmer la valeur de celui-ci . Plusieurs méthodes de détection de changement proposées dans la littérature utilisent ce type de seuil. Parmi ces derniers, on peut citer WILLSKY et JONES (1976) qui propose une méthode de détection de changement dans les systèmes linéaires, et un travail plus récent mené par CHO et collab. (2016) qui propose une méthode basée sur les tests CUSUM pour la détection des points de changement dans un panel de données.

Dans le cas des séries temporelles présentant un comportement dynamique au fil du temps, cette approche peut s'avérer inadéquate. On montre la contrainte de cette approche à l'aide d'un exemple. La figure 4.4 montre le comportement de la statistique de test calculée à partir des données présentant une évolution dynamique dans le temps. La valeur du seuil est fixée dans ce cas à $h = 5$. On peut remarquer que le seuil sélectionné permet de détecter un changement à l'instant $\tau = 100$. En revanche, l'utilisation du même seuil sur la période restante ne permet pas de détecter un changement avec une amplitude plus faible de la statistique de test.

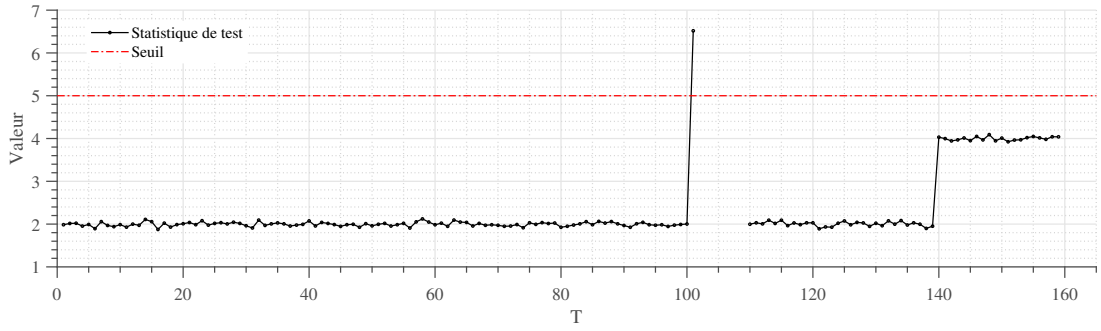


FIGURE 4.4 – Détection de changement en utilisant un seuil fixe. Ce graphique montre les valeurs de statistique de test (noir) et le seuil estimé (rouge).

Cette approche n'est pas adaptée à notre cas, puisque les habitudes de consommation adoptées par les consommateurs montrent une évolution dynamique dans le temps. Ce qui pourrait conduire à des valeurs de statistique de test avec une variance élevée.

4.2.1.2 Seuil adaptatif

Cette approche vise à adapter la valeur de seuil en fonction de la distribution des données au fil du temps. Pour avoir une estimation de la valeur de seuil, nous avons effectué des simulations de type Monte Carlo CHEIFETZ (2013). À partir d'une séquence de données initiales ne présentant pas de changement (comportement normal), un modèle de Markov non homogène est d'abord estimé par la méthode du maximum de vraisemblance. Ensuite, plusieurs séquences (m jeux de données distincts) sont générées à partir de ce modèle et la statistique de test est évaluée pour chacune de ces séquences (voir figure 4.5). Le fractile ($\mathcal{Q}_{1-p}$ avec $0 < p < 1$) de la distribution des statistiques de test est considéré comme la valeur du seuil. Le seuil est estimé de nouveau après chaque détection.

Suivant cette approche, la valeur du seuil est donc déduite directement des données. Elle permet ainsi d'adapter la valeur du seuil aux différents types de changement et de réduire le taux de fausses alarmes. La figure 4.6 montre un exemple de la détection de changement en utilisant un seuil adaptatif. Les seuils estimés sont affichés par des pontillés rouges. Un seuil est estimé au début de la séquence et après chaque détection. Ce qui conduit à des valeurs de seuil $h_1 = 5$ et $h_2 = 3$. Cela permet de détecter les deux changements caractérisés par des amplitudes différentes de la statistique de test. Dans la suite de ce chapitre, cette approche a été adoptée pour déterminer les valeurs de seuil.

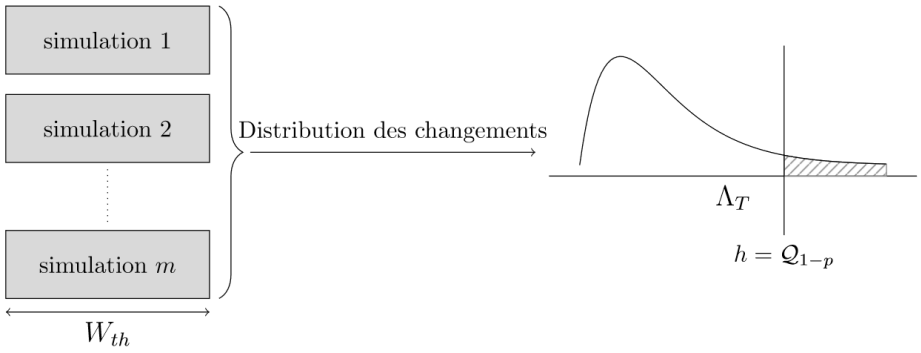


FIGURE 4.5 – Procédure d'estimation du seuil. W_{th} désigne la taille de la fenêtre du seuil et $\mathcal{Q}_{1-p}$ est le fractile de la distribution des statistiques de test considéré comme la valeur du seuil.

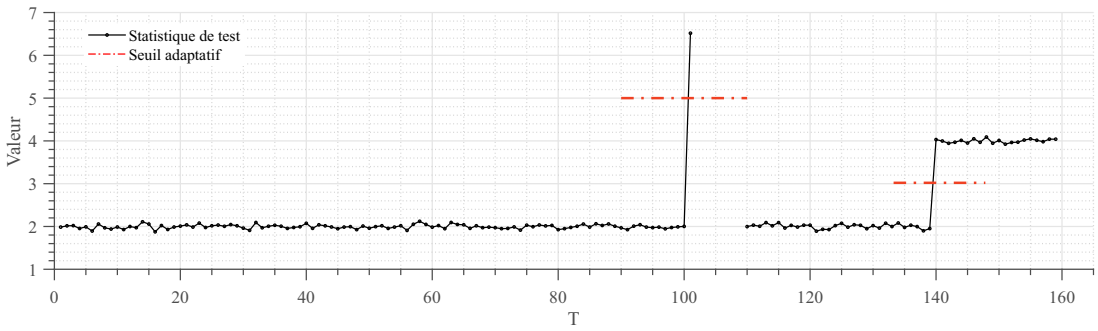


FIGURE 4.6 – Détection de changement en utilisant un seuil adaptatif. Ce graphique montre les valeurs de statistique de test (noir) et les seuils adaptatifs estimés (rouge).

Parmi les recherches effectués pour déterminer un seuil adaptatif, on peut citer l'approche proposée par WANG et collab. (2017) qui utilise les noyaux gaussiennes pour déterminer un intervalle de confiance basé sur l'écart-type de la distribution. Ce dernier est calculé à partir des données au fur et à mesure que celles-ci parviennent.

4.2.2 Détection de changement en ligne

La méthode en ligne permet de détecter de multiples points de changement communs à un ensemble de séquences catégorielles. En intégrant la méthode de détection de changement proposée et le seuil adaptatif décrit ci-dessus dans une plateforme en ligne, la figure 4.7 montre le schéma associé. Dans ce schéma les séquences catégorielles sont représentées en fonction du temps et en utilisant la couleur grise. Les pas de temps correspondant à chaque étape de la méthode en ligne sont affichés en-dessus de la figure (W_{th} : la taille de la fenêtre du seuil, t_0 : la taille initiale de la fenêtre de détection, $\hat{\tau}$: l'instant présentant un changement, T : l'instant où la statistique de test Λ_T dépasse le seuil h). Les fenêtres d'estimation du seuil (rectangles hachurés) et de détection (rectangles en noir) sont affichées en-dessous de l'axe du temps. Dans ce schéma, on distingue deux types de comportement : un comportement jusqu'à l'instant $\hat{\tau}$ (affiché en gris clair) et un autre à partir de $\hat{\tau}$ (affiché en gris foncé).

La procédure en ligne de la détection de changement consiste dans un premier temps à estimer un seuil h. Pour ce faire, la méthode adoptée dans la section 4.2.1.2 est appliquée dans un intervalle de temps désigné par la fenêtre du seuil. Ensuite, la statistique de test est calculée (voir équation (4.26)) en utilisant une fenêtre de taille croissante (avec une taille initiale t_0). Un point de changement est détecté $\hat{\tau}$, lorsque Λ_T dépasse le seuil estimé h. La procédure se réinitialise au niveau de ce dernier. Cette démarche est répétée au fur et à mesure que les données parviennent. Le pseudo code 4 résume les différentes étapes da la méthode de détection de changement en ligne.

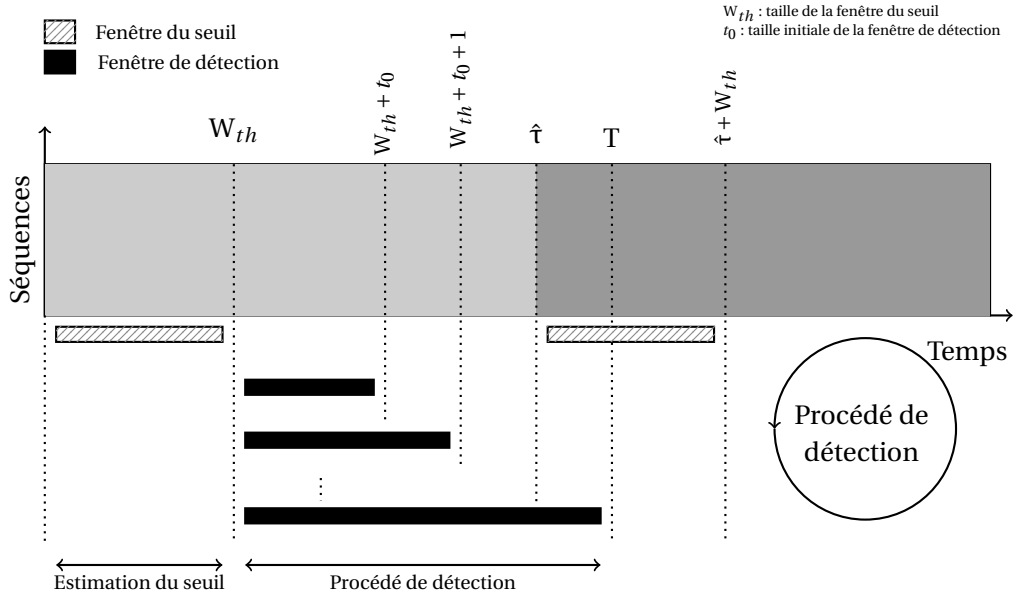


FIGURE 4.7 – Schéma de la méthode en ligne proposée pour la détection des points de changement communs à un ensemble de séquences catégorielles. Deux comportements différents sont affichées : un comportement jusqu'à l'instant $\hat{\tau}$ (en gris clair) et un autre comportement à partir de l'instant $\hat{\tau}$ (en gris foncé). Les différents pas de temps sont notés en-dessus de la figure.

Algorithme 4 : Détection en ligne de changement communs

Données : $Z = (z_1, \ldots, z_T)$, $(\mathbf{u}_1, \ldots, \mathbf{u}_T)$.
Résultat : vecteur des points de changement $\widehat{\boldsymbol{\tau}}$;
initialisation :
fixer la taille de la fenêtre du seuil W_{th};
fixer la taille initiale de la fenêtre de détection t_0;
fixer l'ordre du quantile p;
$a, i, j \leftarrow 0$.

tant que ¬ *fin de la séquence* **faire**
  générer aléatoirement m séquences sur l'intervalle $[a, a+W_{th}]$;
  pour $i = 1, \ldots, m$ **faire**
    $cp(i) \leftarrow \Lambda^{(i)}_{a+W_{th}}(\tau)$ en utilisant Eq. (4.26);
  fin
  $h \leftarrow \mathcal{Q}_{1-p}(cp)$;
  tant que ¬ $(changement)$ **faire**
    $\psi \leftarrow \max_{(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1), \tilde{\boldsymbol{\theta}}_0} \Lambda^{a+W_{th}+t_0+j}_{a+W_{th}}(\tau)$;
    si $\psi > h$ **alors**
      $changement \leftarrow true$;
      $\hat{\tau}(i) \leftarrow \arg\max_\tau \psi$;
      $a \leftarrow \hat{\tau}(i)$;
      $i \leftarrow i+1$;
    sinon
      $j \leftarrow j+1$;
    fin
  fin
fin

retourner $\widehat{\boldsymbol{\tau}}$;

4.3 Étude expérimentale

Dans cette section, dans un premier temps, nous évaluons les performances de la méthode proposée en utilisant des bases de données synthétiques. Une comparaison avec les méthodes de l'état de l'art est également effectuée. Dans un second temps, cette méthode est appliquée sur les bases de données réelles obtenues dans le chapitre précédent (voir figure3.17b et 3.18b).

4.3.1 Méthodes évaluées

Les performances de la méthode de détection de changement proposée sont comparées à celles des méthodes suivantes :

- Modèle de Markov homogène (MM) : ce modèle considère que les données avant et après les points de changement sont distribuées suivant un modèle de Markov homogène d'ordre un. Les variables exogènes n'influencent guère la dynamique d'évolution des séquences ; or, la régularité temporelle des séquences est bien prise en compte ;
- Modèle de régression logistique (LR) : ce modèle considère que les données avant et après les points de changement suivent une loi multinomiale. Les variables exogènes sont utilisées comme variable d'entrée pour ce modèle ; en revanche, la dépendance des états n'est pas prise en compte (GOMBAY et collab., 2017) ;
- Modèle de Markov non-homogène (NMM) : ce modèle considère que les données avant et après les points de changement sont distribuées suivant un modèle de Markov non-homogène. Il tient compte de la régularité temporelle dans l'évolution d'états et de facteurs exogènes ayant une influence sur cette évolution.

Ces trois méthodes ont été adaptées au traitement en ligne de séquences catégorielles. Pour avoir une comparaison équitable des méthodes, les paramètres liés à la détection en ligne (tailles des fenêtres de seuil et de détection) sont initialisés avec les mêmes valeurs pour l'ensemble des méthodes.

L'ensemble des variables de contexte introduites dans la section 3.2.3 du chapitre précédent sont également utilisées comme variable d'entrée pour les modèles mentionnés.

4.3.2 Extension des critères d'évaluation au cas multi-segment

Les critères d'évaluation décrits dans la suite sont basés sur la matrice de confusion. Les matrices de confusion sont généralement utilisées dans le contexte d'une classification binaire. Dans le cas d'un point de changement, les segments avant et après le changement sont binarisés (les observations sont remplacées par 0 avant le changement et par 1 après le changement). Ainsi, le comportement normal est indiqué par des zéros et le changement de comportement par des uns. Dans le cadre de multiples points de changement, cette procédure est répétée en considérant un point de changement à la fois. Ainsi, nous obtenons une matrice de confusion par point de changement. Cette matrice fournit quatre types d'information concernant les détections effectuées :

- *faux positifs* (FP) : le taux de comportements normaux (0) identifiés par erreur comme de changements de comportement (1),
- *faux négatifs* (FN) : le taux de changements (1) qui n'ont pas été détectés,
- *vrais positifs* (VP) : le taux de changements (1) correctement identifiés,
- *vrais négatifs* (VN) : le taux de comportements normaux (0) correctement identifiés.

Les critères suivants sont utilisés pour évaluer la performance des méthodes mentionnées :

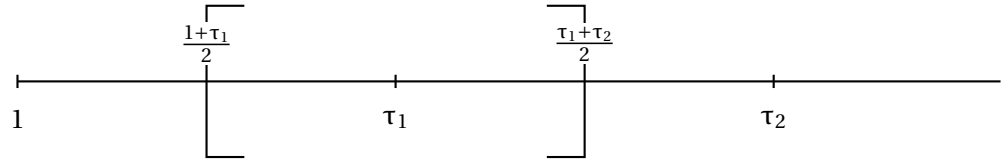
- F-mesure : combinant la précision et la sensibilité dans un seul critère, ce critère est le plus souvent utilisé dans le cadre des méthodes de détection de changements CAO et collab. (2015) ;

- Aire sous la courbe ROC (AUC) : mesure la performance globale d'une méthode de détection selon différentes valeurs de seuil. Ce critère peut atteindre une valeur maximale de 1 ; les valeurs proches de 1 indiquent la bonne performance des méthode évaluées ;
- Taux de Fausses alarmes (FA) : ce critère est le même que le taux *faux positifs* (décrit ci-dessus) et indique le taux de comportement normaux identifiés par erreur comme de changements de comportement ;
- Taux de vrais positifs (TPR) : est le taux de changements de comportement correctement identifiés. Dans le cas de la détection de multiples points de changement, ce critère est moyenné sur l'ensemble de ces derniers ;
- Délai de détection moyen (DD) : est le délai nécessaire pour une méthode à détecter un changement. Dans le cas d'une détection de multiples points de changement, ce critère est moyenné sur l'ensemble des délais.

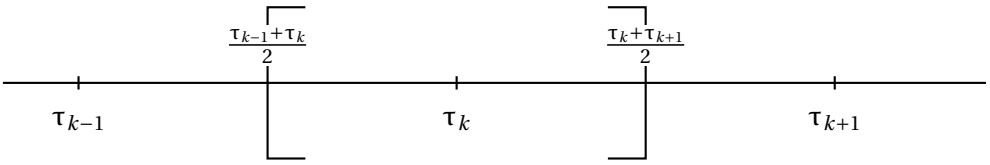
L'ensemble des critères mentionnés ci-dessus (à l'exception du délai de détection) sont généralement utilisés dans le cadre d'une classification binaire (deux segments). Toutefois, les méthodes de détection de changement introduites sont adaptées pour détecter de multiples points de changement dans plusieurs séquences catégorielles. Une extension des critères d'évaluation pour le cas multi-segments est proposée dans la suite.

Étapes permettant de calculer les critères d'évaluation

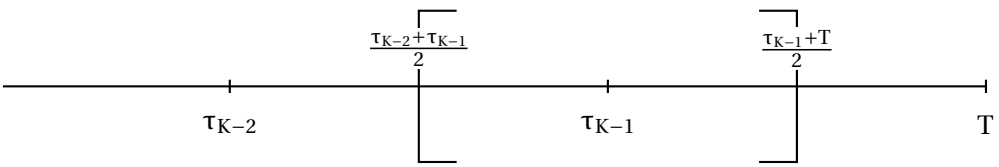
1. Cette étape consiste à associer les points de changement détectés ($\hat{\tau}$) aux vrais points de changement (τ). Pour ce faire, on définit des intervalles de recherche autour de chaque point de changement. Selon la position temporelle du point de changement dans la séquence examinée, trois scénarios différents sont envisageables (voir figure 4.8). Dans ces figures, on suppose que $\tau_0 = 1$ et $\tau_K = T$, où K désigne le nombre de segments.



(a) Cas du premier point de changement



(b) Cas d'un point de changement intermédiaire



(c) Cas du dernier point de changement

FIGURE 4.8 – Intervalle de recherche d'un point de changement

2. Cette étape a pour objectif de calculer les critères d'évaluation mentionnés pour chaque point de changement associé lors de la première étape. On décrit cette étape à l'aide d'un exemple (voir figure 4.9). Dans cette figure, la première séquence indique les vrais points

de changement, et la deuxième présente ceux estimés par un algorithme de détection. On s'intéresse au point de changement τ_2. Suivant l'étape 1 (deuxième scénario), on associe $\hat{\tau}_3$ à τ_2 (point de changement estimés le plus proche). Ensuite, les segments se trouvant avant et après ces derniers dans l'intervalle $[\frac{\tau_1+\tau_2}{2}, \frac{\tau_2+\tau_3}{2}]$ sont remplacés respectivement par 0 et par 1. Cela permet de calculer les critères d'évaluation pour ce point de changement. Cette étape est répétée pour chaque vrai point de changement.

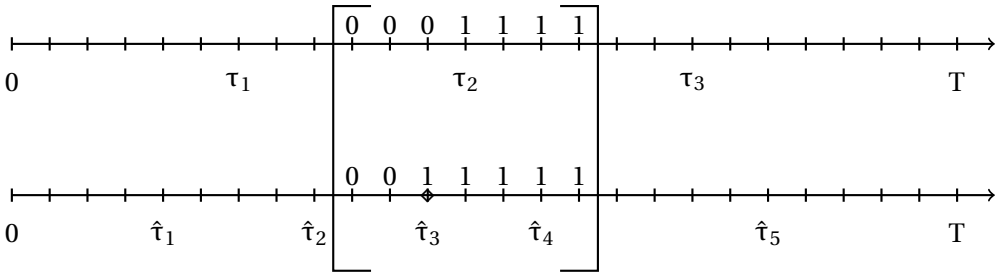


FIGURE 4.9 – Exemple illustratif utilisé pour l'étape 2 de la méthode d'évaluation

En suivant cette méthode, on obtient une valeur par point de changement. Pour calculer les critères d'évaluation sur la séquence entière, on utilise la formule suivante :

$$\mathrm{E} = \frac{1}{\mathrm{C}} \sum_{i=1}^{\mathrm{C}} \mathrm{E}_{\tau_i} \tag{4.29}$$

où E désigne un des critères d'évaluation mentionnés et C indique le nombre de vrais points de changement.

4.3.3 Cas d'étude : données simulées

Cette section examine les performances, en termes de différents critères d'évaluation, des méthodes de détection mentionnées sur les bases de données synthétiques. La procédure adoptée pour générer ces bases de données, les résultats obtenus, ainsi qu'une comparaison entre les méthodes mentionnées sont décrites dans la suite.

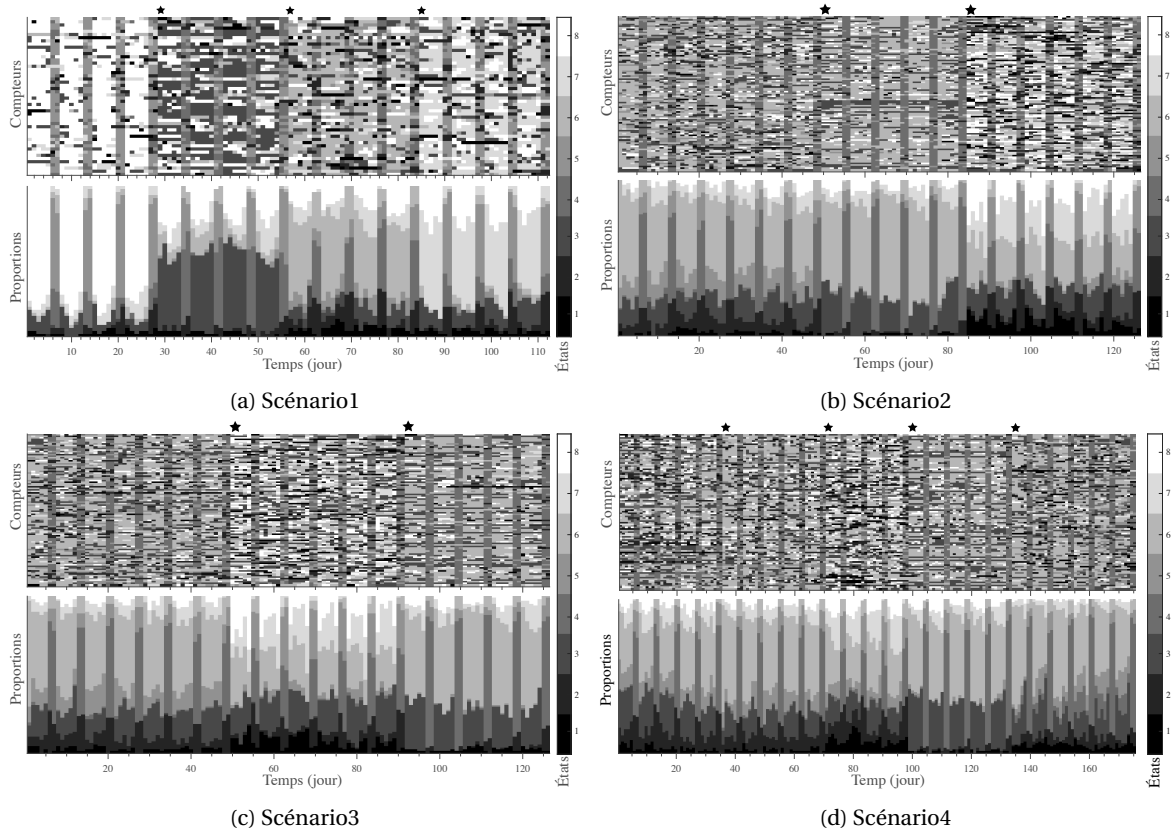
4.3.3.1 Procédure de simulation des séquences

À partir des paramètres estimés d'un modèle de Markov non-homogène sur des habitudes journalières de consommation issues de données réelles, 4 scénarios différents ont été conçus (voir tableau 4.1).

TABLEAU 4.1 – Bases de données simulées. 4 scénarios sont conçus en utilisant un modèle de Markov non-homogène. La paire (n,T) désigne le nombre de séquences et le nombre d'instants ; l'ensemble des instants présentant un vrai point de changement est indiqué par une liste ; la détectabilité est exprimée par des barres dont la longueur est proportionnelle à la hétérogénéité des segments avant et après les points de changement. Une barre plus longue désigne une hétérogénéité plus importante des segments.

Base de données	(n,T)	Points de changement	Détectabilité
Scénario1	(50,112)	[29,57,85]	
Scénario2	(100,126)	[50,85]	
Scénario3	(100,126)	[50,85]	
Scénario4	(100,175)	[36,71,99,134]	

Dans ce tableau le nombre de séquences et le nombre d'instants sont indiqués par la paire (n, T) et l'ensemble des instants présentant un changement est est indiqué par une liste. Dans ce tableau, on propose également une mesure de détectabilité, qui est la norme L_2 entre les paramètres



(a) Scénario1

(b) Scénario2

(c) Scénario3

(d) Scénario4

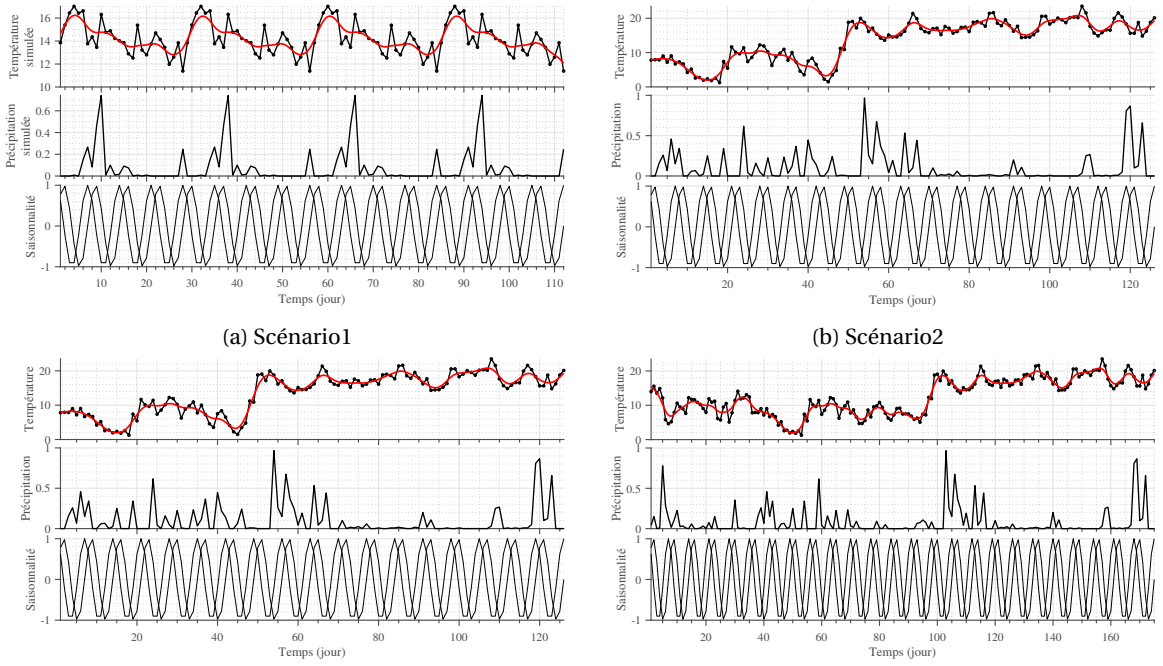
FIGURE 4.10 – Visualisation des bases de données synthétiques correspondant aux 4 scénarios détaillés dans le tableau 4.1. Chaque figure est constituée de deux types de graphiques : le premier montre l'évolution des catégories au fil du temps et le deuxième montre l'évolution de la proportion des catégories. Les vrais points de changements sont indiqués en utilisant des étoiles en-dessus de chaque figure.

du modèle de Markov avant et après les points de changement. Cette mesure est moyennée sur l'ensemble des points de changement et exprimée par des barres dont la longueur est proportionnelle à la hétérogénéité des segments avant et après les points de changement. Une valeur élevée de cette mesure (une barre plus longue) indique une hétérogénéité plus importante des segments et ainsi une détection plus facile des changements.

Les bases de données correspondant aux scénarios mentionnés sont également visualisées dans la figure 4.10. Pour chaque scénario, deux types de graphiques sont utilisés. Les graphiques du haut présentent les bases de données générées sur lesquelles les méthodes de détection de changement sont appliquées, et les graphiques du bas présentent la proportion des catégories au fil du temps. Les changements sont indiqués par des étoiles en dessus de chaque figure et les états sont représentés par de plusieurs niveaux de gris (K = 8 états). Les graphiques présentant la proportion des états permettent d'avoir une vue macroscopique sur les vrais points de changement et pourraient faciliter l'interprétation des résultats obtenus.

En regardant ces graphiques (proportion des états), on peut remarquer une différence dans la proportion des états avant et après les changements. Le « Scénario1 » (voir figure 4.10a) contient trois points de changement. La différence dans la proportion des états est assez nette avant et après le premier point de changement pour ce scénario ; cette différence est moins prononcée concernant les segments avant et après le troisième point de changement.

Le « Scénario2 » (voir figure 4.10b) contient deux points de changement. Ce dernier est caractérisé par une évolution de la proportion d'états très similaire sur toute la période de simulation. Concernant « Scénario3 » (voir figure 4.10c) qui inclut également deux points de changement, la différence dans la proportion des états avant et après les points de changement est plus facilement repérable. La mesure de détectabilité est également plus élevée pour ce scénario en comparaison



(a) Scénario1

(b) Scénario2

(c) Scénario3

(d) Scénario4

FIGURE 4.11 – Variables d'entrée correspondant aux bases de données affichées dans la figure 4.10.

des autres scénarios conçus. Ainsi, la différence entre les paramètres estimés avant et après les points de changement pour ce scénario est plus prononcée. Concernant le « Scénario4 » (voir figure 4.10d), les changements ne peuvent pas être facilement repérés, ce qui explique également le niveau de détectabilité plus faible pour ce scénario.

Le vecteur des variables d'entrée **u** associé à chacun des scénarios inclut la température (en °C), le niveau de précipitation (en mm) et la saisonnalité présentée par des termes trigonométriques. Ces variables sont visualisées dans la figure 4.11. Concernant les variables associées au premier scénario (voir figure 4.11a), on remarque un motif répétitif des facteurs climatiques. Les données de ce scénario sont générées à partir de 4 dynamiques markoviennes estimées sur la même période. Les figures 4.11a et 4.11c montrent respectivement les variables associées aux scénarios 2 et 3. On note un changement de la moyenne dans la courbe de température à partir de l'instant $t = 45$ (moyenne plus élevée). Concernant le « Scénario4 » (voir figure 4.11d), on remarque la même hausse de température que dans les scénarios précédant à partir de l'instant $t = 100$. La courbe de précipitation montre une variabilité irrégulière pour l'ensemble des scénarios. Finalement, le graphique montrant la saisonnalité des états correspond à deux des termes trigonométriques pour les périodes considérées. Ces derniers sont ceux qui ont été introduits dans la section 3.2.3.

4.3.3.2 Détection des points de changement pour les séquences simulées

La méthode de détection de changements proposée a été appliquée aux quatre scénarios mentionnés dans la section précédente. À titre d'illustration, les résultats obtenus sur les séquences du « Scénario2 » sont affichés dans la figure 4.12.

La statistique de test Λ_T est calculée de manière séquentielle sur les séquences catégorielles du deuxième scénario. Le seuil adaptatif est estimé sur des périodes déterminées par la fenêtre de seuil (flèches en pointillés en dessous de l'axe des abscisses) et affiché par les pointillés horizontaux sur le graphique. On note que les valeurs de la statistique de test dépassent 2 fois les seuils estimés. Les points de changement détectés peuvent être repérés à l'aide des croix sur l'axe des abscisses. La fenêtre de détection (taille initiale) est également indiquée en dessous de l'axe des

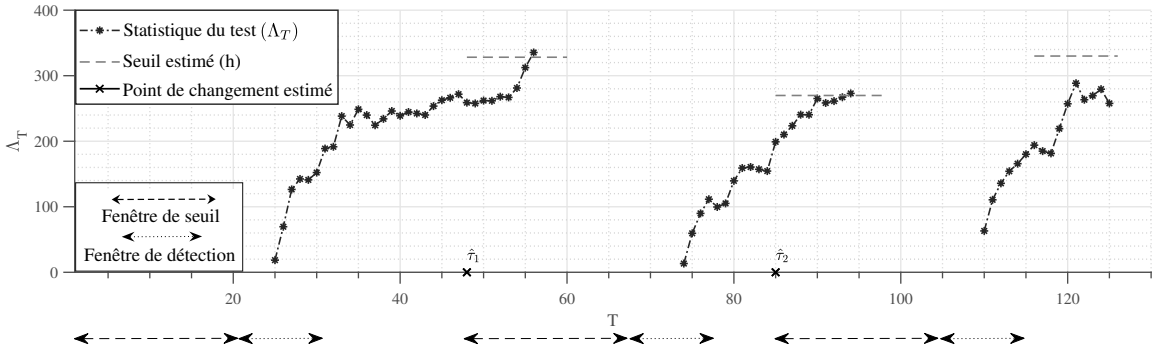


FIGURE 4.12 – Statistique de test calculée à l'aide de la méthode de détection de changement proposée sur les séquences du « Scénario2 ». La fenêtre dédiée à l'estimation du seuil adaptatif et la taille initiale de la fenêtre de détection sont affichées en dessous de l'axe des abscisses.

abscisses.

La méthode de détection de changement proposée commence par estimer une valeur de seuil en utilisant $m = 20$ bases de séquences générées de taille $W_{th} = 21$ instants. On suppose que cette période ne présente pas de changements. Le fractile d'ordre $1 - p = 0.05$ de la distribution des points de changement estimés sur l'ensemble des m séquences fournit une première valeur de seuil. La statistique de test est ensuite calculée en utilisant une fenêtre de taille croissante dont la taille initiale est $t0 = 14$. Cette dernière dépasse la valeur de seuil pour la première fois à l'instant $t = 56$. La méthode proposée s'interrompe et l'instant associé à la valeur maximale de la statistique de test est considéré comme un point de changement $\hat{\tau}_1 = 48$. La durée entre l'instant de la détection et le vrai point de changement constitue le délai de détection.

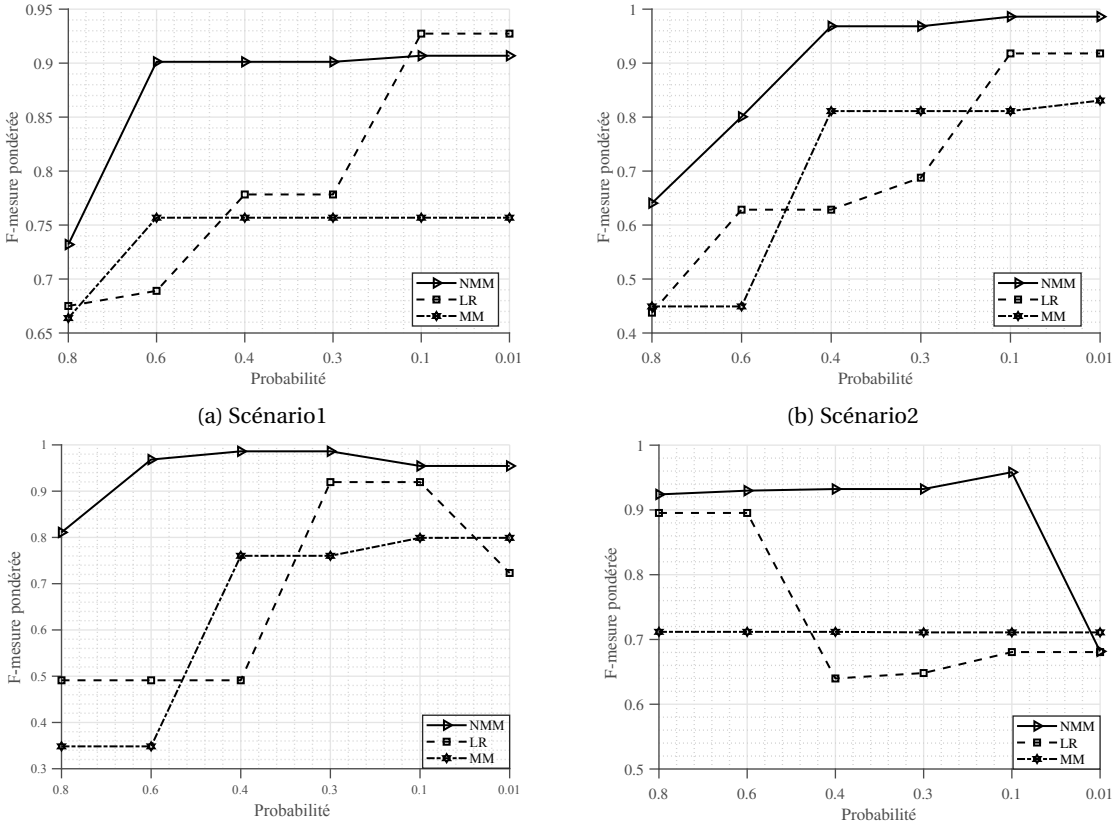
La procédure est réinitialisée au niveau du dernier point de changement détecté. En suivant la même démarche, un nouveau seuil est estimé et la statistique de test est calculée sur la période restante. Ainsi, un nouveau point de changement est estimé à l'instant $\hat{\tau}_2 = 85$. Le calcul de la statistique de test sur le dernier segment ne conduit pas à une nouvelle détection. En regardant la figure 4.10b, on peut noter que les points de changement détectés sont en accord avec les points de changements réels.

La section suivante compare la performance des méthodes de détection de changement introduites en utilisant les différents critères d'évaluation.

4.3.3.3 Comparaison des méthodes évaluées

Pour comparer la performance des méthodes mentionnées sur les quatre bases de données synthétiques, nous avons utilisé des outils graphiques et numériques basés sur les critères d'évaluation proposés. Dans un premier temps, le critère de F-mesure proposé est calculé en fonction de différentes valeurs de probabilité p (voir figure 4.13). La probabilité p indique l'ordre du fractile ($\mathcal{Q}_{1-p}$) de la distribution des points de changement qui est utilisée lors de l'estimation d'un seuil adaptatif. Une valeur faible de p correspond à un seuil plus élevé. Dans la suite, une comparaison basée sur d'autres critères d'évaluation (aire sous la courbe ROC, taux de vrais positifs et délai de détection) est également effectuée (voir tableau 4.2). Ces critères sont moyennés sur l'ensemble des valeurs de p. Concernant les critères AUC et TPR, une valeur proche de 1 indique une bonne performance de la méthode utilisée. Concernant le délai de détection (DD), une valeur plus faible (délai plus court) est meilleure. Les meilleures performances obtenues sur chacune des bases de données sont marquées en gras.

En regardant la figure 4.13, on remarque une bonne performance de la méthode proposée (NMM) en termes de la F-mesure pour l'ensemble des scénarios. Concernant le premier scénario (voir figure 4.13a), on note une F-mesure légèrement plus élevée en utilisant la méthode LR



(a) Scénario1

(b) Scénario2

(c) Scénario3

(d) Scénario4

FIGURE 4.13 – Évaluation des méthodes en termes de F-mesure pour les quatre scénarios conçus

pour les faibles valeurs de p (seuil plus grand). Ce scénario présente une mesure de détectabilité moyenne (voir tableau 4.1). Toutefois, la méthode proposée reste assez proche et fournit la meilleure performance au regard de l'ensemble des critères utilisés (voir tableau 4.2). La performance de la méthode basée sur le modèle de Markov homogène (MM) reste inférieure pour ce scénario.

Concernant les scénarios 2 et 3 (voir figures 4.13b et 4.13c), la méthode proposée fournit la meilleure performance, peu importe la valeur du seuil. La méthode LR donne une meilleure performance par rapport à la méthode MM pour les valeurs de p élevées.

Concernant le « Scénario4 » (voir figure 4.13d) qui présente une détectabilité plus faible par rapport aux autres scénarios, la méthode MM fournit une F-mesure constante. Cela s'explique par le fait que la dynamique des transitions reste relativement homogène sur toute la période (voir figure 4.10d). Ainsi, la détection de changement devient très compliquée en utilisant cette méthode qui ne tient pas compte des variables de contexte. La méthode LR montre de bonnes performances en utilisant $p = 0.8$ et $p = 0.6$. Pour les valeurs faibles de p, c'est-à-dire les valeurs élevées du seuil, la méthode LR ne permet pas de détecter l'ensemble des points de changements. La meilleure performance est obtenue en utilisant la méthode proposée NMM et $p = 0.1$. L'un des points de changement reste non identifié en utilisant la méthode proposé avec $p = 0.01$, ce qui explique la chute observée pour cette valeur.

Le tableau 4.2 fournit le délai de détection pour chaque méthode. En utilisant la méthode LR, le délai de détection est légèrement plus court pour les scénarios 1, 2 et 4. Toutefois, cette méthode est moins performant que la méthode proposée en termes des autres critères (AUC et TPR). La méthode MM présente un délai de détection plus court pour le scénario 4, mais ses performances en termes des critères AUC et TPR sont moins élevées que les autres méthodes.La méthode proposée présente le délai de détection le plus court pour le scénario 3 et donne les meilleures performances

au regard de l'ensemble des critères d'évaluation.

TABLEAU 4.2 – Tableau de comparaison. Les méthodes évaluées sont basées sur les modèles suivants : modèle de Markov homogène (MM), modèle de régression logistique (LR), modèle de Markov non-homogène (NMM). Les critères d'évaluation sont : aire sous la courbe ROC (AUC), taux de vrais positifs (TPR), délai de détection (DD). Ces critères sont moyennés sur les différentes valeurs de seuil.

Model	**Metric**	**Scénario1**	**Scénario2**	**Scénario3**	**Scénario4**
MM	AUC	0.84	0.77	0.80	0.70
	TPR	0.74	0.78	0.62	0.65
	DD	6.6	7.5	5	**2.6**
LR	AUC	0.86	0.82	0.81	0.83
	TPR	0.84	0.73	0.65	0.76
	DD	**4**	**5**	8	6.3
NMM	AUC	**0.89**	**0.92**	**0.95**	**0.91**
	TPR	**0.94**	**0.96**	**0.97**	**0.96**
	DD	4.6	7.5	**4**	8.7

Finalement, le tableau 4.3 fournit, pour chacun des scénarios, le nombre de fausses alarmes en fonction de différentes valeurs de probabilité p. Ce tableau peut contenir trois types de valeurs :

- Valeur positive : le nombre de fausses alarmes ;
- Zéro : les détections correspondent aux vrais changements ;
- Valeur négative : le nombre de changement non identifié.

On note un nombre élevé de fausses alarmes pour les méthodes MM et LR. En regardant ces tableaux, on peut noter un lien entre l'ordre du fractile p et le nombre de fausses alarmes (valeur positive) ou le nombre de changements non identifiés (valeur négative). Les valeurs plus élevées de p (les valeurs plus faibles de seuil) conduisent à un nombre élevé de fausses alarmes. Alors que les valeurs plus faibles de p conduisent à un nombre croissant de changements non identifiés. Ainsi, le choix de p est primordial pour la méthode de détection de changement.

TABLEAU 4.3 – Nombre de faussses alarmes en fonction des valeurs de p pour chaque scénario. Dans ces tableaux, une valeur positive désigne le nombre de fausses alarmes, une valeur négative désigne le nombre de changements non identifiés et le zéro indique la correspondance entre les détections et les vrais points de changement.

(a) Scénario1

p	MM	LR	NMM
0.8	3	2	1
0.6	1	2	0
0.4	1	1	0
0.3	1	1	0
0.1	1	0	0
0.01	1	0	0

(b) Scénario2

p	MM	LR	NMM
0.8	2	2	1
0.6	1	2	0
0.4	1	2	0
0.3	1	2	0
0.1	1	0	0
0.01	1	0	0

(c) Scénario3

p	MM	LR	NMM
0.8	3	3	1
0.6	3	3	0
0.4	2	3	0
0.3	2	1	0
0.1	1	1	0
0.01	1	2	0

(d) Scénario4

p	MM	LR	NMM
0.8	1	0	0
0.6	1	0	0
0.4	1	-1	0
0.3	1	0	0
0.1	1	-1	0
0.01	1	-1	-1

4.3.4 Cas d'étude : réseau d'eau potable

Dans cette partie, les résultats de l'application de la méthode de détection de changement proposée sur de multiples séquences catégorielles issues du réseau d'eau sont présentés. Les expérimentations menées ont été réalisées au sein des groupes homogènes issus de la classification préalable de compteurs (voir figure 3.17b et 3.18b). Pour ces bases de données, nous ne disposons pas d'information supplémentaire concernant les vrais points de changement. Pour vérifier les points de changement obtenus, nous avons analysé la distribution (à l'aide de la méthode d'estimation par noyau (SIMONOFF, 2012)) des points de changement estimés au niveau de chaque compteur issu du groupe examiné. L'objectif est de vérifier si certains pics de cette distribution sont en accord avec les changements détectés en travaillant sur l'ensemble de compteurs. Le tableau 4.4 décrit les deux groupes de compteurs analysés.

TABLEAU 4.4 – Caractéristiques des groupes de compteurs analysés et liste des points de changement obtenus

Base	Groupe	Pas de temps (t)	# compteurs	Figure	Durée (T)	Points de changement
1	3	Journalier	206	4.14	251 jours	[50, 104, 167, 203, 237]
2	13	Hebdomadaire	267	4.16	78 semaines	[10, 22, 36, 48, 59]

Base 1 : Groupe 3 dans le cas d'états journaliers

Ce groupe est issu de la classification de séquences d'états journaliers (voir figure 4.14a). Ce dernier est constitué d'un ensemble de 206 compteurs pour lesquels nous avons extrait les habitudes de consommation durant une période de 251 jours (du 16 mars 2015 au 21 novembre 2015). La figure 4.14b représente l'évolution de la proportion des états au fil du temps. Les périodes de vacances scolaires et d'été sont également encadrées par des pointillés rouges. On note une hausse de la proportion de l'état vert (état 6) pendant les périodes de vacances et une hausse de la proportion de l'état 2 pendant les vacances d'été 2015. En dehors des périodes de vacances, on remarque que l'état orange (état 7) fait partie des états dominants du groupe 3. Cet état est caractérisé par un seul pic dans l'après-midi qui représente le profil « administratif ».

La méthode de détection de changement proposée a été appliquée à l'ensemble des compteurs appartenant à ce groupe et la figure 4.14c montre le calcul de la statistique de test Λ_T. Le seuil adaptatif est estimé sur une fenêtre de $W_{th} = 21$ jours et les valeurs obtenues sont représentées par des pointillés rouges sur ce graphique. On peut remarquer les différentes valeurs de seuil qui sont estimées après chaque détection. La fenêtre de détection, qui est une fenêtre de taille croissante, est initialisée à $t_0 = 14$ jours. Les points de changement détectés sont affichés à l'aide de croix sur l'axe des abscisses.

Au total, 5 points de changement ont été détectés sur la période examinée. Les changements détectés correspondent respectivement à la fin de la période de vacances scolaires du mois d'avril, au début et à la fin de la période d'été 2015, au début des vacances scolaires du mois d'octobre et à une nouvelle habitude adoptée après les vacances du mois d'octobre. D'après les résultats obtenus, on constate que les changements de nature plus ponctuelle (portant sur une courte durée) correspondent à des valeurs faibles de la statistique de test ($\hat{\tau}_1$, $\hat{\tau}_4$ et $\hat{\tau}_5$). En revanche, les changements persistants dans le temps sont associés à des valeurs plus grandes de la statistique de test. Parmi ceux-ci, on peux citer les points de changement $\hat{\tau}_2$ et $\hat{\tau}_3$ qui sont associés au début et à la fin d'été.

La distribution des points de changement estimés au niveau de chaque compteur est également tracée dans la figure 4.14d. Les pics les plus importants dans cette densité correspondent à des points de changement détectés en travaillant sur l'ensemble de compteurs. Cela pourrait confirmer les points de changements estimés en absence de vrais labels.

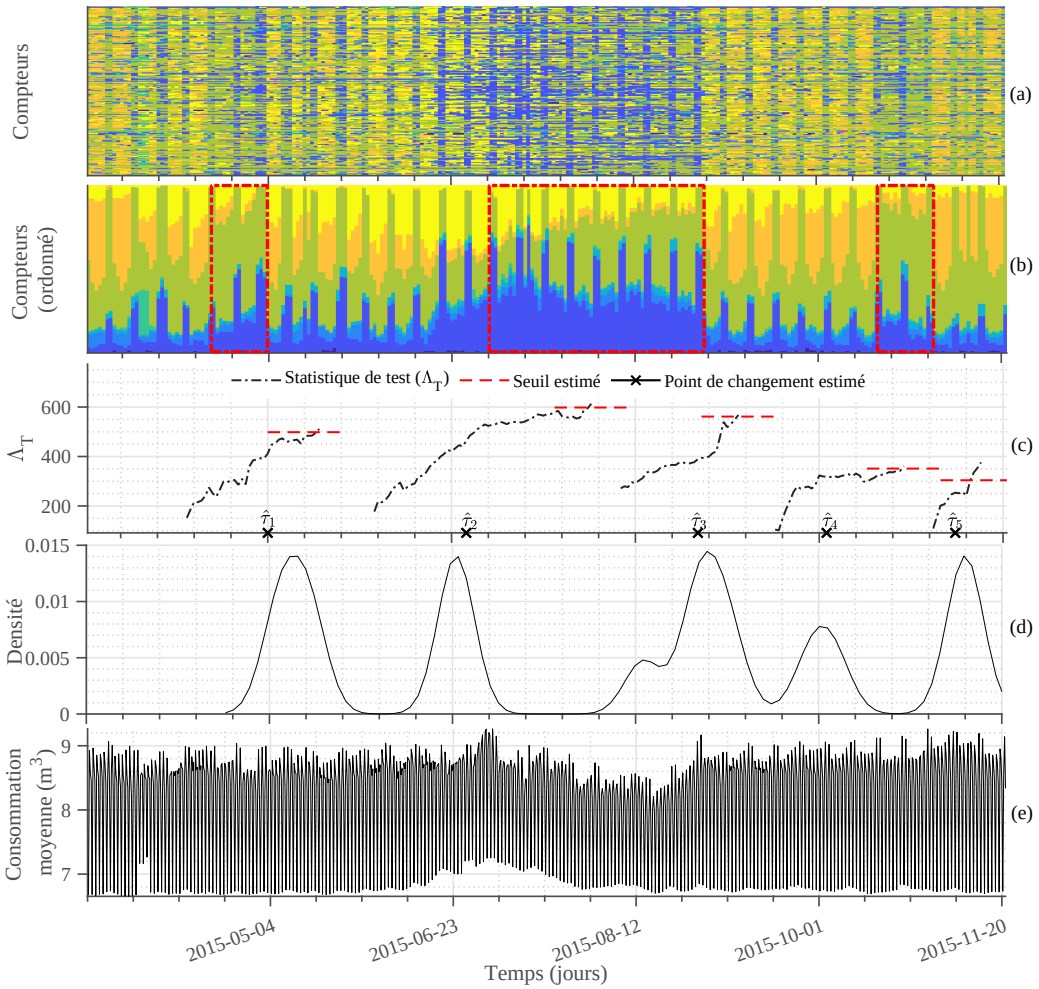


FIGURE 4.14 – Détection de changement par la méthode proposée, qui est appliquée aux séquences du groupe 3 dans le cas d'états journaliers ; (a) 206 séquences présentant l'évolution d'habitudes journalières de consommation durant 251 jours, chaque ligne correspond à un compteur ; (b) évolution de la proportion des états ; (c) statistique de test et points de changement détectés ; (d) densité des points de changement obtenus au niveau de chaque compteur ; (e) consommation moyenne du groupe en mètres-cubes (24 relevés par jour).

Nous avons également affiché la courbe de consommation (en mètres-cubes) moyennée sur l'ensemble de compteurs (voir figure 4.14e). On peut noter un changement à partir du mois de juin (point de changement $\hat{\tau}_2$) où la consommation moyenne croît progressivement et revient au même niveau constaté avant l'été (point de changement $\hat{\tau}_3$). La section suivante fournit une analyse plus approfondie de l'influence des variables climatiques et des évènements calendaires.

Analyse de l'influence des variables exogènes

On s'intéresse ici à l'analyse de l'influence des variables de contexte sur le deuxième point de changement détecté ($\hat{\tau}_2 = 104$). Le tableau 4.5 fournit les coefficients estimés du modèle pour les transitions les plus significatives avant et après ce changement.

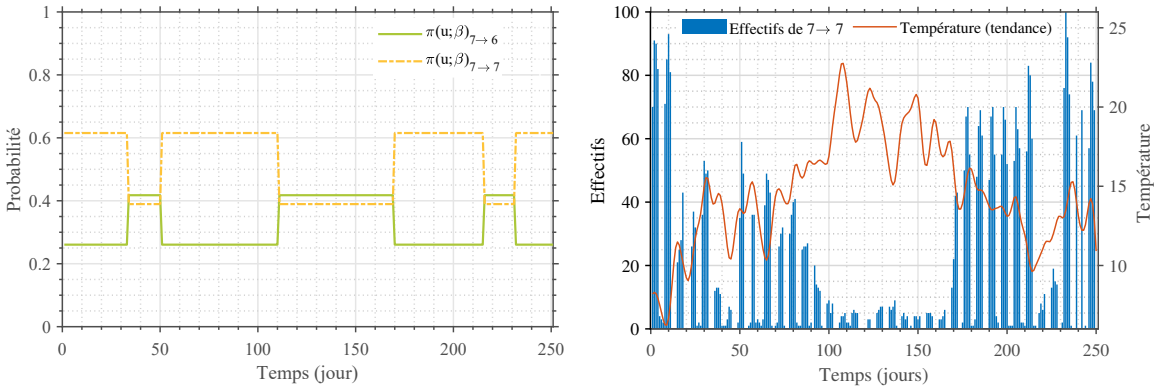
On peut noter que les coefficients de la variable calendrier valent zéro pour l'ensemble des transitions avant le changement. Cela est dû au fait que cette période ne comporte pas d'évènements calendaires (vacances scolaires ou jours fériés). On constate une présence significative de la transition de l'état 2 vers lui-même pendant la période après le changement (29%). Les coefficients de la température et de la précipitation sont en opposition avec ceux estimés sur le premier segment. Cela confirme une hausse de température et une baisse de précipitation après ce chan-

TABLEAU 4.5 – Coefficient estimé du modèle correspondant aux segments avant et après le deuxième point de changement $\hat{\tau}_2 = 104$

Transitions	$t < \hat{\tau}_2$				$t > \hat{\tau}_2$			
$(s_{t-1} \rightarrow s_t)$	Proportion	T	P	C	Proportion	T	P	C
$2 \rightarrow 2$	0,10	-2,63	40,99	0	0,29	0,15	-3,14	-1,26
$2 \rightarrow 6$	0,03	-2,63	39,49	0	0,06	-0,02	-2,80	-1,50
$2 \rightarrow 8$	0,03	-2,71	40,03	0	0,04	0,30	-2,90	-1,31
$6 \rightarrow 2$	0,02	0,67	0,46	0	0,05	-0,45	7,27	-7,62
$6 \rightarrow 6$	0,25	0,71	0,81	0	0,22	-0,63	7,61	-8,88
$6 \rightarrow 7$	0,03	0,55	0,85	0	0,01	-0,73	7,51	-9,27
$6 \rightarrow 8$	0,04	0,72	1,26	0	0,02	-0,54	6,90	-8,83
$7 \rightarrow 6$	0,04	-0,12	0,14	0	0,01	0,01	-7,50	1,87
$7 \rightarrow 7$	0,08	-0,18	0,61	0	0,01	-0,06	-5,56	0,61
$8 \rightarrow 2$	0,02	0,28	-2,57	0	0,04	0,84	2,91	0,92
$8 \rightarrow 6$	0,03	0,28	-0,98	0	0,03	0,62	4,85	1,30
$8 \rightarrow 8$	0,12	0,33	-0,75	0	0,07	0,55	4,86	0,34

gement (début de la période d'été). Les transitions de l'état 7 vers les états 6 et 7 et de l'état 8 vers lui-même ont une proportion plus élevée pendant la période avant le changement. On remarque les coefficients plus faibles de la température pendant cette période, ce qui indique l'influence de la température sur ces transitions. Les coefficients de la précipitation ont également les valeurs les plus élevées (négatives) pour les transitions de l'état 7 après le changement. Ainsi, un niveau de précipitation très faible pendant la période d'été pourrait être à l'origine de ces transitions. Ce changement est caractérisé par un changement de saison. Cette analyse peut également être menée sur les autres points de changement détectés.

Dans la figure 4.15, nous avons analysé l'influence des évènements calendaires et de la température sur toute la période pour les transitions de l'état 7. L'évolution des probabilités de transition $\pi(u;\beta)_{k,\ell}$ (calculées à partir des paramètres estimés du modèle β pour la variable d'entrée u « évènements calendaires ») de l'état 7 vers les états 6 et 7 est affichée dans la figure 4.15a. On remarque une probabilité faible de maintenir la même habitude (état 7 est associé à « activité professionnelle ») pendant les périodes de vacances scolaires. Les probabilités de transition de l'état 7 vers l'état 6 sont plus élevées pendant ces périodes. Cette analyse montre également la capacité de la méthode à modéliser la dépendance entre les transitions et les variables exogènes.



(a) Probabilités de transition (calendrier)

(b) Influence de température

FIGURE 4.15 – Analyse de l'impact des variables exogènes ; (a) analyse des probabilités de transition de l'état 7 vers les états 6 et 7 en fonction des évènements calendaires ; (b) analyse de l'impact de la variable température sur l'évolution des effectifs de la transition de l'état 7 vers lui même

La figure 4.15b montre l'évolution des effectifs de transition de l'état 7 vers lui-même (à l'aide

d'un diagramme en barres) au fil du temps. La tendance de la température est également superposée sur ce graphique. On constate que ces deux dernières sont liées par une relation opposée. Lorsque la température croît, l'effectif de cette transition diminue. Cela justifie l'influence de cette variable sur les transitions et ainsi sa prise en compte par le modèle.

Base 2 : Groupe 13 dans le cas d'états hebdomadaires

Les compteurs analysés dans cette partie sont issus du groupe 13 de la classification de séquences d'états hebdomadaires (voir figure 3.18b). Ce groupe est constitué d'un ensemble de 267 compteurs et son évolution d'habitudes de consommation couvre une période de 78 semaines du 26 janvier 2015 au 24 juillet 2016 (voir figure 4.16a).

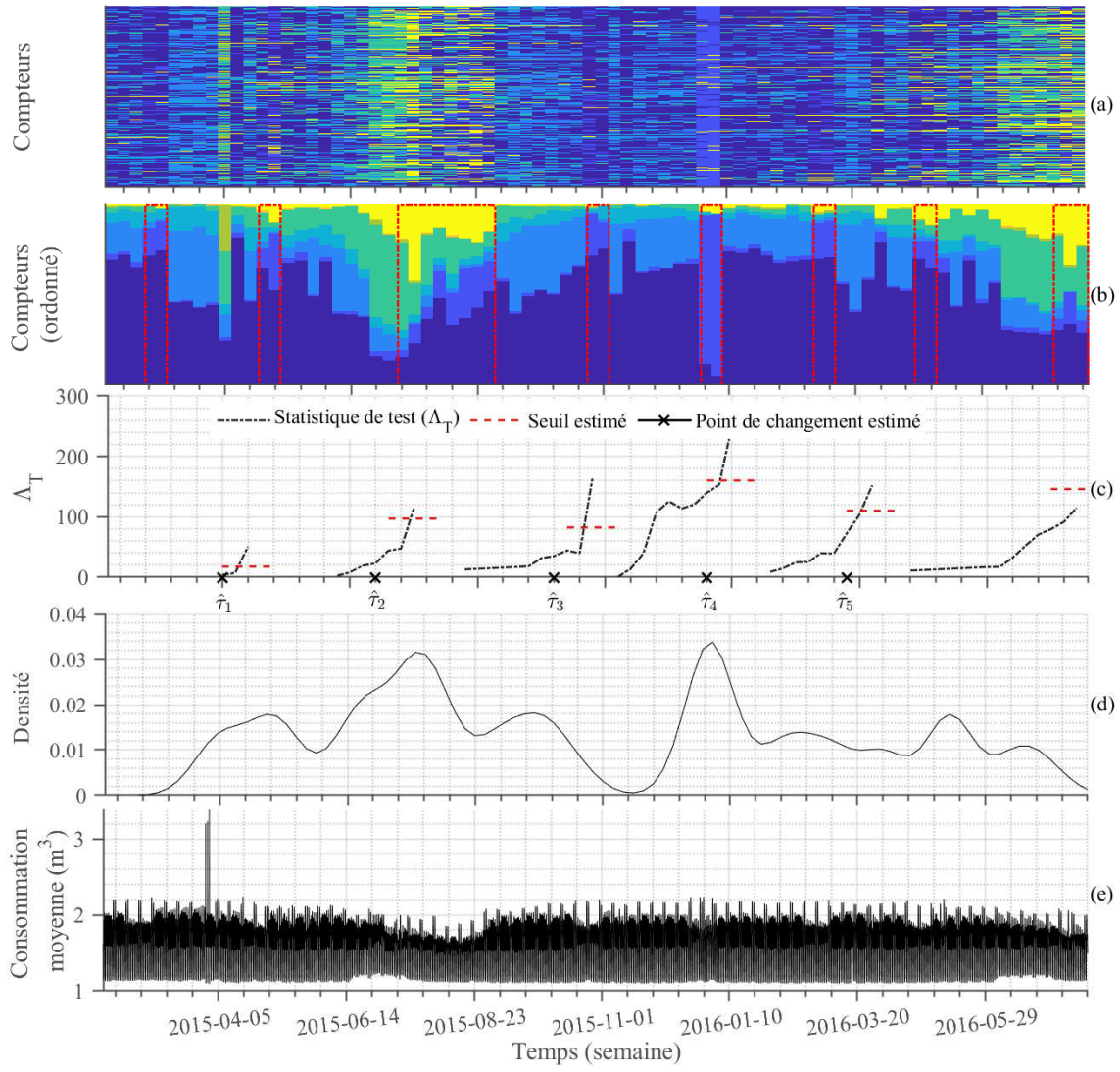


FIGURE 4.16 – Détection de changement par la méthode proposée, qui est appliquée aux séquences du groupe 13 dans le cas d'états hebdomadaires ; (a) 267 séquences présentant l'évolution d'habitudes hebdomadaires de consommation durant 78 semaines, chaque ligne correspond à un compteur ; (b) évolution de la proportion des états ; (c) statistique de test et points de changement détectés ; (d) densité des points de changement estimés au niveau de chaque compteur ; (e) consommation moyenne du groupe en mètres-cubes (168 relevés par semaine)

Les graphiques présentés sont les mêmes que ceux utilisés précédemment. En regardant la proportion des états au fil du temps (voir figure 4.16b), on constate que les changements dans les habitudes de consommation des compteurs sont liés le plus souvent à des évènements ca-

lendaires (vacances scolaires et périodes d'été encadrées par des pointillés rouges). En regardant simultanément la courbe de consommation moyenne pour ce groupe (voir figure 4.16e), on note un pic de consommation aux environs du mois d'avril qui correspond à une augmentation de la proportion de l'état vert pendant cette période (voir figure 4.16a). On note un autre changement dans la consommation (une baisse) à partir du mois de juin qui persiste jusqu'à la fin du mois d'août (la période d'été). Ce dernier correspond à une hausse de la proportion de l'état 8 (jaune) pendant cette période.

La méthode de détection de changement proposée a été appliquée sur les séquences de ce groupe et la figure 4.16 (c) montre le calcul de la statistique de test. La taille de la fenêtre de seuil considérée est $W_{th} = 7$ semaines et la taille initiale de la fenêtre de détection est fixée à $t_0 = 2$ semaines. Les points de changements détectés sont affichés à l'aide de croix sur l'axe des abscisses. La statistique de test calculée sur la dernière fenêtre ne dépasse pas le seuil estimé. Au total, 5 points de changement sont détectés.

La distribution des points de changement estimés au niveau de chaque compteur est visualisée dans la figure 4.16d. À l'exception du dernier point de changement, les pics les plus importants dans la densité correspondent à des points de changement détectés en travaillant sur l'ensemble de compteurs. On peut observer que le dernier changement ($\hat{\tau}_5$) a été détecté plus tôt en appliquant la méthode proposée sur l'ensemble de compteurs.

Le premier point de changement détecté $\hat{\tau}_1 = 10$ (05/04/2015) est lié à une anomalie dans les habitudes de consommation à partir duquel on note un changement dans le comportement des compteurs. Cette période est influencée également par une hausse progressive de la température (mois de juin et juillet). Le deuxième changement détecté est associé au début de la période estivale (été 2015) où on remarque un changement dans la proportion des états présents pendant cette période. Le troisième changement détecté correspond au retour au comportement normal avant l'été. Le quatrième point de changement désigne la période de vacances de Noël où l'état 2 devient l'état dominant pendant cette période. Finalement, le dernier point de changement détecté correspond à un changement de comportement progressif jusqu'à l'été 2016.

Analyse de l'influence des variables exogènes

Nous nous focalisons ici sur l'analyse de l'impact des variables exogènes sur le premier point de changement détecté $\hat{\tau}_1 = 10$. Le tableau 4.6 fournit les coefficients estimés du modèle concernant les transitions les plus significatives avant et après ce changement. On constate une proportion très importante des transitions de l'état 1 vers lui-même; cette proportion étant plus faible après le changement. Le coefficient de la température pour cette transition est plus élevé (négatif) après le changement. Ainsi, la hausse dans la température conduit à une diminution de la proportion de cette transition après le changement (habitude adoptée par la catégorie « personnes actives »). On peut également observer le même motif concernant des transitions de l'état 1 vers l'état 3. Le coefficient de précipitation associé à la transition de l'état 5 vers l'état 1 présente une valeur positive avant le changement et une valeur négative après ce dernier. Ainsi, une baisse de précipitation pourrait être à l'origine de l'augmentation de la proportion de cette transition après le changement. Finalement, on observe une proportion plus élevée des transitions de l'état 5 vers lui-même après le changement. Les coefficients de précipitation et de température associés ont des signes (+/-) opposés avant et après le changement. Cela traduit le fait qu'une hausse de température accompagnée par une baisse de précipitation pourraient accroître la proportion de cette transition après le changement. L'état 5 est caractérisé par un pic du soir légèrement plus élevé par rapport à celui du matin et ce dernier est plus décalé dans le temps.

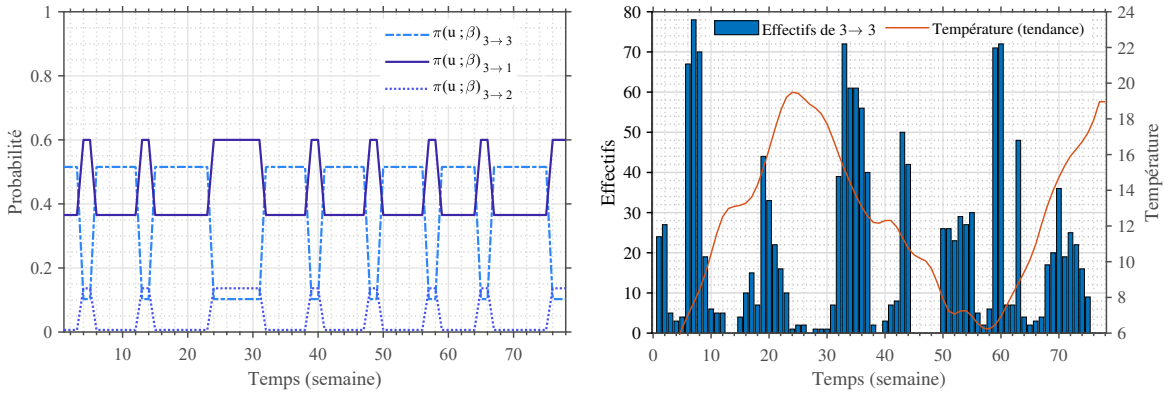
L'influence des évènements calendaires et de la température est également étudiée sur toute la période pour les transitions de l'état 3 (voir figure 4.17). La figure 4.17a montre l'évolution des probabilités de transition $\pi(u;\beta)_{k,\ell}$ de l'état 3 vers les états 1, 2 et 3 pour la variable « évènements calendaires ». On peut remarquer que la probabilité de maintenir la même habitude de consomma-

TABLEAU 4.6 – Coefficient estimé du modèle (β) correspondant aux segments avant et après le premier point de changement $\hat{\tau}_1 = 10$

Transitions	$t < \hat{\tau}_2$				$t > \hat{\tau}_2$			
$(s_{t-1} \rightarrow s_t)$	Proportion	T	P	C	Proportion	T	P	C
$1 \rightarrow 1$	0,44	0,06	-6,58	-8,87	0,35	-0,16	-4,31	-1,71
$1 \rightarrow 3$	0,10	0,43	-5,92	-11,14	0,08	-0,02	-3,16	-3,52
$1 \rightarrow 5$	0,01	-0,05	-2,54	-11,21	0,05	0,12	-2,52	-2,36
$3 \rightarrow 1$	0,07	0,17	12,41	6,55	0,07	0,05	-4,05	-1,21
$3 \rightarrow 3$	0,13	0,40	-13,69	5,15	0,04	0,48	-3,31	-3,16
$4 \rightarrow 4$	0,04	-0,01	-13,83	-1,81	0,02	-0,27	-3,58	-0,72
$5 \rightarrow 1$	0,01	-0,85	11,57	-3,90	0,06	-0,46	-2,61	-0,91
$5 \rightarrow 5$	0,01	-0,22	10,47	3,95	0,04	0,25	-2,30	-1,55

tion diminue pendant les vacances scolaires, et les consommateurs ayant adopté cette habitude (état 3) ont tendance de changer leur comportement et d'exercer une nouvelle habitude (états 1 et 2).

Nous avons également analysé l'évolution des effectifs des transitions de l'état 3 vers lui-même en fonction de la température (voir figure 4.17b). Ce graphique met en évidence que la présence de cette transition est liée à une baisse dans la température. Cela justifie la prise en compte de cette variable dans le modèle.



(a) Probabilités de transition (calendrier)

(b) Influence de température

FIGURE 4.17 – Analyse de l'impact des variables exogènes ; (a) analyse des probabilités de transition de l'état 3 vers les états 1 et 3 en fonction des évènements calendaires ; (b) analyse de l'impacte de la variable température sur l'évolution des effectifs de la transition de l'état 3 vers lui-même

4.4 Conclusion

Ce chapitre s'est focalisé sur la détection de changements communs à un ensemble de séquences catégorielles non-stationnaires. La non-stationnarité des données implique que la répartition des catégories ainsi que les transitions entre celles-ci peuvent évoluer dans le temps. Pour modéliser cette évolution, nous avons proposée une méthode de détection de changement basée sur les tests séquentiels du rapport de vraisemblance généralisé, qui est fondée elle-même sur les modèles de Markov non-homogènes. Ce choix a été motivé par le comportement dynamique des séquences d'habitudes de consommation et par leurs dépendances à des facteurs exogènes. Ce modèle permet de modéliser cette évolution conjointe de la dynamique des habitudes au fil du temps. Pour pouvoir détecter différents types de changement, vu la non stationnarité des données, un seuil adaptatif est utilisé. Ce seuil est calculé à partir des données observées et est mis à jour après chaque détection.

Les expérimentations menées sur les données synthétiques ont montré les bonnes performances de la méthode proposée. Celle-ci a été comparée avec deux autres méthodes, basées l'une sur le modèle de Markov homogène et l'autre sur le modèle de régression logistique. Cette méthode a également fait l'objet d'une application sur les données réelles issues d'une classification préalable de compteurs obtenue dans le chapitre précédent. Les résultats obtenus se sont révélés cohérents au regard des éléments contextuels disponibles (changement de saisons, variation de la courbe température, variations du volume d'eau consommé, vacances scolaires). L'analyse de l'influence des variables de contexte, en utilisant les coefficients estimés du modèle, a permis d'interpréter et de caractériser les changements détectés.

Chapitre 5

Conclusion et perspectives

Synthèse

Le problème étudié dans le cadre de cette thèse porte dans un premier temps sur la classification des séquences catégorielles de nature non stationnaire. Plus concrètement, on s'est focalisé sur le regroupement des consommateurs d'eau potable dont le comportement est dynamique et évolue dans le temps. Il est à noter qu'une étape préliminaire de discrétisation des profils saisonniers (journaliers ou hebdomadaires) a permis d'extraire de chaque compteur, l'ensemble de ses habitudes de consommation (voir le chapitre 2). Pour modéliser la dynamique des habitudes de consommation, un mélange de modèles de Markov non homogènes est proposé dans le chapitre 3. Cette méthode s'est révélée pertinente pour modéliser la dynamique conjointe des habitudes de consommation ; chaque habitude de consommation à chaque pas de temps dépend d'un nombre de facteurs exogènes. Les expérimentations menées sur les données simulées ont également confirmé l'efficacité du modèle proposé pour le regroupement des séquences temporelles qui dépendent des variables exogènes.

Pour estimer les paramètres du modèle de mélange proposé, l'algorithme CEM a été adopté. Une fois les paramètres du modèle estimés, ceux-ci ont été utilisés pour prédire les futures habitudes au sein de chaque groupe de compteurs. Les comparaisons des méthodes, en termes d'erreurs de prévision, ont montré les bonnes performances du modèle proposé pour la prévision, à un pas de temps, de futurs états.

Dans un second temps, nous nous sommes intéressés au problème de la détection de changements communs à un ensemble de séquences catégorielles. Pour résoudre ce problème, dans le chapitre 4, nous nous sommes appuyés sur les tests statistiques séquentiels du rapport de vraisemblance. La particularité de cette méthode réside dans l'hypothèse que les séquences avant et après les points de changement sont distribuées suivant des modèles de Markov non homogènes.

Pour identifier les changements structurels de nature différente, un seuil adaptatif a été utilisé. Ce dernier est estimé à l'aide des simulations Monte Carlo sur des périodes de données ne présentant pas de changement. L'évaluation de cette méthode sur les données simulées a mis en évidence la pertinence de ce choix.

Par ailleurs, la méthode de détection de changement proposée a été appliquée sur les séquences d'habitudes de consommation issues d'un réseau d'eau potable. Les changements détectés se sont avérés pertinents à l'égard des variables contextuelles. Finalement, les coefficients estimés du modèle ont permis de caractériser et d'interpréter les changements détectés.

Perspectives

Dans cette thèse, nous avons proposé des outils pour analyser le comportement des usagers. Il serait ainsi intéressant d'étendre les méthodes proposées et d'approfondir les analyses. Dans la suite, nous envisageons quelques orientations de recherche qui méritent d'être explorées.

Nous nous sommes focalisés, dans cette étude, sur l'analyse du comportement journalier et hebdomadaire des consommateurs. Cette analyse pourrait également être effectuée sur les tranches temporelles plus fines. Par exemple, on peut imaginer deux habitudes de consommation par jour, celle du matin et celle du soir. Cela permettrait de distinguer les consommateurs selon leur comportement routinier dans la journée. La méthode SAX (*symbolic aggregate approximation*) (LIN et collab., 2003) peut être utilisée pour discrétiser les périodes de tailles différentes.

Nous avons considéré plusieurs variables de contexte pour la modélisation et l'interprétation des résultats, à savoir les variables climatiques, les évènements calendaires, le volume de consommation et la périodicité. Les résultats obtenus pourraient être confirmés à l'aide des variables démographiques et socioéconomiques qui donnent une information concernant la catégorie socioprofessionnelle des consommateurs. Cela nécessite l'acquisition des données supplémentaires. Il fait également partie des perspectives de mener une analyse spatiale sur un ensemble de compteurs dont on connaît leur catégorie socioprofessionnelle. Ainsi, les liens spatiaux des compteurs ayant adopté des habitudes de consommations similaires peuvent être étudiés.

Pour le regroupement des séquences d'habitudes de consommation, nous nous sommes appuyés sur un mélange de modèles de Markov non homogènes. Pour l'estimation des paramètres de ce modèle, l'algorithme CEM a été exploité. À court terme, il serait intéressant de comparer le regroupement obtenu à l'aide de cet algorithme avec celui de l'algorithme EM. Une autre extension de ce modèle peut porter sur le traitement séquentiel des séquences catégorielles. Dans ce dernier, le regroupement est mis à jour au fur et à mesure que les données parviennent. Pour ce faire, on peut se baser sur une version en ligne de l'algorithme EM qui est proposée par SAMÉ et collab. (2007) et SATO et ISHII (2000).

Dans cette étude, le regroupement statique des séquences catégorielles a été étudié. Ce dernier implique que l'affectation des séquences aux groupes ne change pas dans le temps. Ensuite, en s'appuyant sur les tests séquentiels du rapport de vraisemblance, chaque groupe a fait l'objet d'une segmentation temporelle de l'ensemble de ses séquences. Par ailleurs, nous avons observé que les habitudes de consommation sont sujettes à des changements persistants dans le temps, ce qui traduit le comportement dynamique de celles-ci. Dans ce cadre, on peut considérer une classification dynamique des séquences catégorielles, où les classes peuvent évoluer dans le temps. Pour effectuer cette classification dynamique, on introduit ci-dessous quelques extensions envisagées :

- le modèle de l'analyse des transitions latentes (LTA) (COLLINS et LANZA, 2009) peut être utilisé pour modéliser l'évolution des classes des compteurs dans le temps. Ainsi, la classification estimée initialement peut évoluer suivant les probabilités de transition. Les transitions entre les classes peuvent être considérées comme des changements de comportement communs à un groupe de compteurs. On peut également imaginer une extension séquentielle de cette méthode, où les classes sont mises à jour au fur et à mesure que les données parviennent ;

- l'approche de classification par bloc (appelé *Co-clustering*) proposée par GOVAERT et NADIF (2003) peut être utilisée pour une classification simultanée des observations et des variables. Dans ce contexte, NADIF et GOVAERT (2005) propose un modèle de mélange par bloc où les variables latentes associées au regroupement des individus et de leurs caractéristiques sont estimées à l'aide de l'algorithme EM. Une extension de ce modèle dans le cas des séquences catégorielles est proposée par GAY et collab. (2015). Pour tenir compte de la dépendance temporelle des séquences, ce modèle impose une contrainte sur la classification des seg-

ments dans le temps. Une extension de ce modèle serait de modéliser chaque bloc à l'aide d'un modèle de Markov non-homogène, ce qui permet également de tenir compte des facteurs exogènes.

Annexe A

Estimation des paramètres

L'équation (4.22) peut s'écrire par une combinaison de trois fonctions de vraisemblance, chacune liée à un vecteur de paramètres différent :

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_{0\ell}) + \mathcal{L}(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_{1\ell}) - \mathcal{L}(\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\beta}}_{0\ell}). \tag{A.1}$$

Les paramètres optimaux $\boldsymbol{\theta}^{opt} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_\ell)$ peuvent être estimés séparément pour chaque fonction de vraisemblance, en maximisant :

$$\mathcal{L}(\boldsymbol{\theta}^{opt}) = \log \mathrm{P}(\mathbf{z}_1, \ldots, \mathbf{z}_\mathrm{T}; \boldsymbol{\theta}^{opt}) = \log \mathrm{P}(\mathbf{z}_1; \boldsymbol{\theta}^{opt}) + \sum_{t=2}^{\mathrm{T}} \log \mathrm{P}(\mathbf{z}_t \mid \mathbf{z}_{t-1}; \boldsymbol{\theta}^{opt}). \tag{A.2}$$

En développant l'équation (A.2), on obtient :

$$\mathcal{L}(\boldsymbol{\theta}^{opt}) = \mathcal{L}_1(\boldsymbol{\alpha}) + \sum_{\ell=1}^{\mathrm{K}} \mathcal{L}_{2,\ell}(\boldsymbol{\beta}_\ell), \tag{A.3}$$

où

$$\mathcal{L}_1(\boldsymbol{\alpha}) = \sum_{\ell=1}^{\mathrm{K}} \sum_{i=1}^{n} z_{i1\ell} \log \pi_\ell(\mathbf{u}_{i1}; \boldsymbol{\alpha}) \tag{A.4}$$

et

$$\mathcal{L}_{2,\ell}(\boldsymbol{\beta}_\ell) = \sum_{t=2}^{\mathrm{T}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n} z_{i(t-1)\ell}\, z_{itk} \log \pi_{k\ell}(\mathbf{u}_{it}; \boldsymbol{\beta}_\ell). \tag{A.5}$$

Afin de renforcer la robustesse dans l'estimation des paramètres du modèle de Markov, un terme de régularisation a été considéré. Ainsi, on maximise :

$$\underset{\boldsymbol{\theta}^{opt}}{\arg\max}\, (\mathcal{L}(\boldsymbol{\theta}^{opt}) - \frac{\lambda}{2} \|\boldsymbol{\theta}^{opt}\|_2^2), \tag{A.6}$$

où $\|.\|_2$ désigne la norme L_2 et λ est un hyper-paramètre qui contrôle l'importance du terme de régularisation. Pour tout $\lambda > 0$, la fonction d'optimisation devient strictement convexe et garantie l'obtention d'une solution unique. Pour estimer les paramètres du modèle, nous avons opté pour $\lambda = 1\mathrm{e}{-8}$. Par conséquent, le problème (A.6) est résolu par $\mathrm{K}+1$ problèmes de maximisation :

$$\begin{aligned} &\underset{\boldsymbol{\beta}_\ell}{\arg\max} \left[\mathcal{L}_{2,\ell}(\boldsymbol{\beta}_\ell) - \frac{\lambda}{2} \|\boldsymbol{\beta}_\ell\|_2^2 \right], \quad \forall \ell = 1, \ldots, \mathrm{K} \\ &\underset{\boldsymbol{\alpha}}{\arg\max} \left[\mathcal{L}_1(\boldsymbol{\alpha}) - \frac{\lambda}{2} \|\boldsymbol{\alpha}\|_2^2 \right]. \end{aligned} \tag{A.7}$$

Les paramètres de chaque modèle de Markov sont estimés par la méthode du maximum vraisemblance, qui est mise en oeuvre via l'algorithme Newton Raphson (Roos et collab., 1998) également connu, dans cette situation, sous le nom *Iteratively Reweighted Least Squares* (IRLS) (Holland et Welsch, 1977).

A.1 Expressions du vecteur gradient et de la matrice hessienne

L'hypothèse nécessaire pour pouvoir utiliser l'algorithme d'IRLS est que le problème de maximisation (A.6) doit être différentiable par rapport au vecteur des paramètres $\boldsymbol{\theta}$. La première (vecteur du gradient) et la seconde dérivée partielle (matrice hessienne) de $\mathcal{L}_1$ et $\mathcal{L}_{2,\ell}$ sont définies comme suit :

$$
\begin{aligned}
\frac{\partial \mathcal{L}_1(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_\ell} &= \sum_{i=1}^{n} \Big[(z_{i,1,\ell} - \pi_\ell(\mathbf{u}_{i,1}; \boldsymbol{\alpha}) \Big] \mathbf{u}_{i,1}, \\
\frac{\partial^2 \mathcal{L}_1(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\alpha}'_\ell} &= -\sum_{i=1}^{n} \pi_k(\mathbf{u}_{i,1}; \boldsymbol{\alpha}) \times \Big[\mathbb{1}\{k = \ell\} - \pi_\ell(\mathbf{u}_{i,1}; \boldsymbol{\alpha}) \Big] \mathbf{u}_{i,1} \mathbf{u}'_{i,1}, \\
\frac{\partial \mathcal{L}_{2,\ell}(\boldsymbol{\beta}_\ell)}{\partial \boldsymbol{\beta}_{k,\ell}} &= \sum_{t=2}^{\mathrm{T}} \sum_{i=1}^{n} \Big[(z_{i,t-1,\ell}\, z_{i,t,k} - \pi_{k,\ell}(\mathbf{u}_{i,t}; \boldsymbol{\beta}_\ell) \Big] \mathbf{u}_{i,t}, \\
\frac{\partial^2 \mathcal{L}_{2,\ell}(\boldsymbol{\beta}_\ell)}{\partial \boldsymbol{\beta}_{k,\ell} \partial \boldsymbol{\beta}'_{h,\ell}} &= -\sum_{t=2}^{\mathrm{T}} \sum_{i=1}^{n} \pi_{k,\ell}(\mathbf{u}_{i,t}; \boldsymbol{\beta}_\ell) \times \Big[\mathbb{1}\{k = h\} - \pi_{h,\ell}(\mathbf{u}_{i,t}; \boldsymbol{\beta}_\ell) \Big] \mathbf{u}_{i,t}\, \mathbf{u}'_{i,t}.
\end{aligned}
$$

Annexe B

Classification des séquences présentant des profils manquants

Cet annexe présente une extension de la méthode proposée permettant de classifier de séquences catégorielles, pour regrouper les séquences d'états présentant un nombre d'état manquants au fil du temps. Lors de la phase de prétraitement des relevés de consommation, un grand nombre de compteurs (52% soit 4279 compteurs) présentant des valeurs manquantes ont été exclus des analyses. En effet, les méthodes de prétraitement utilisées imposent l'existence des valeurs de consommation sur au moins 7 jours consécutifs. L'objectif de cette partie est de montrer comment les compteurs exclus ont pu finalement être classifiés. Nous nous focalisons ici sur des séquences d'états journaliers. La méthodologie utilisée s'appuie sur les deux étapes suivantes.

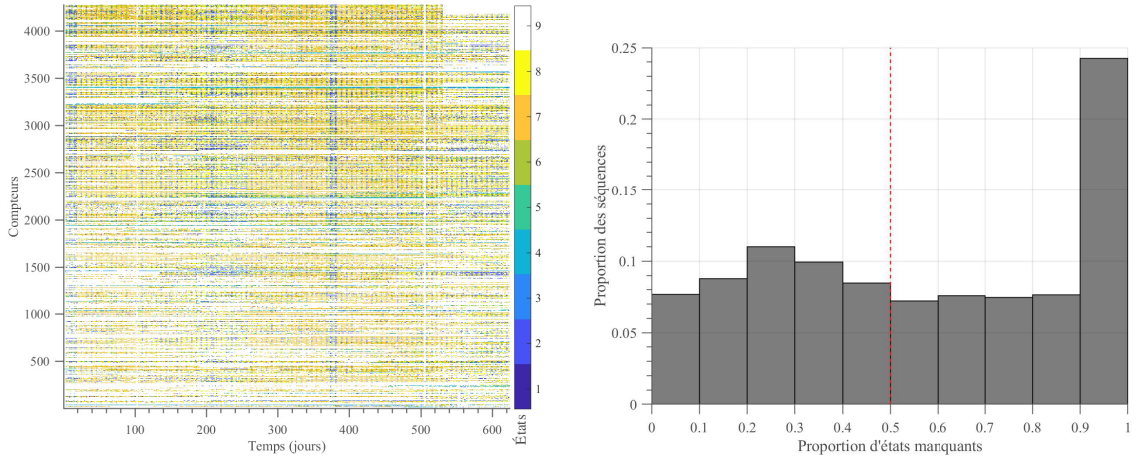
Étape 1 : affectation des profils journaliers aux états

Cette étape n'est réalisée que sur les profils journaliers qui ne comportent pas de valeurs manquantes. Cette étape est réalisée en affectant chaque profil à l'état le plus probable (voir la figure 2.9) a posteriori (règle du maximum a posteriori), en utilisant le modèle FReMix :

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

où $\mathcal{N}(.; \boldsymbol{\mu}; \boldsymbol{\Sigma})$ est la densité gaussienne de vecteur moyen $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$ et $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_k, \alpha_1, \ldots, \alpha_k, \sigma_1^2, \ldots, \sigma_k^2)$ est le vecteur des paramètres du modèle qui sont estimés à partir des données ne présentant pas d'états manquants. Ainsi, la classification des profils journaliers issus des séquences présentant d'états manquants est opérée sur les séries notées $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ avec $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$ où $m = 24$ dans le cas des profils journaliers. À l'issue de cette étape, nous avons obtenu les données présentées dans la figure B.1a où les états manquants sont représentés par la couleur blanche (état 9).

La figure B.1b montre l'histogramme de la proportion d'états manquants par compteur. On peut observer que plus de 24% des compteurs (parmi les 4 279 compteurs) contiennent plus de 90% d'états manquants. Nous avons considéré un seuil (ligne verticale rouge) de 50% d'états manquants par compteur. Les compteurs ayant une proportion d'états manquants inférieur à ce seuil sont éligibles pour l'étape d'affectation (41% des compteurs soit 1 778 compteurs).



(a) (b)

FIGURE B.1 – (a) : Ensemble des 4 279 compteurs contenant de valeurs manquantes. Les valeurs manquantes sont représentées par l'état 9 (couleur blanche) ; (b) : Histogramme de la proportion d'états manquants par compteur.

Étape 2 : affectation des compteurs aux classes

Deux solutions sont possibles : la première consiste simplement à appliquer la règle du maximum a posteriori en considérant que les paramètres du modèle MixJNMM sont déjà connus (Etapes E et C de l'algorithme CEM). La seconde approche qui est plus générale consiste à adapter l'algorithme CEM aux cas de séquences d'état susceptibles de présenter des valeurs manquantes. Cette variante de l'algorithme, qui nécessite par conséquent de ré-estimer les paramètres du modèle, consiste à maximiser la log-vraisemblance complétée suivante à l'aide d'un algorithme CEM analogue à celui déjà présenté :

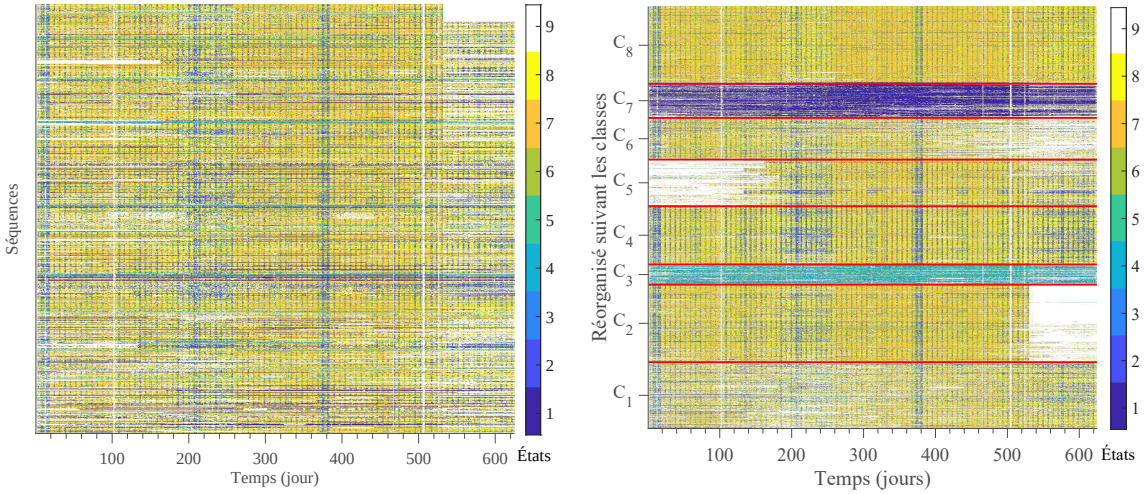
$$\begin{aligned} \mathcal{L}_c^{(m)}(\boldsymbol{\phi}) &= \log \mathrm{P}(\mathbf{z}, \mathbf{w} \mid \mathbf{u}; \boldsymbol{\phi}) \\ &= \sum_{g=1}^{\mathrm{G}} \left[\sum_{i=1}^{n} w_{ig} \log p_g + \sum_{i=1}^{n} \sum_{\ell=1}^{\mathrm{K}} w_{ig} \delta_{i,1} z_{i1\ell} \log \pi_{g\ell}(\mathbf{u}_{i1}; \boldsymbol{\alpha}_g) \right. \\ &+ \left. \sum_{i=1}^{n} \sum_{t=2}^{\mathrm{T}} \sum_{k=1}^{\mathrm{K}} \sum_{\ell=1}^{\mathrm{K}} w_{ig} \delta_{it} \delta_{i(t-1)} z_{itk} z_{i(t-1)\ell} \log \pi_{g\ell k}(\mathbf{u}_{it}; \boldsymbol{\beta}_{g\ell}) \right], \end{aligned} \tag{B.2}$$

où

— $\delta_{i,1} = 1$ si l'état initial de la séquence i existe, sinon $\delta_{i,1} = 0$;

— $\delta_{it}\delta_{i(t-1)} = 1$ si les états aux temps t et $t-1$ existent, sinon $\delta_{it}\delta_{i(t-1)} = 0$.

A l'issue de la seconde étape, et en suivant la deuxième approche, 1 778 compteurs contenant des valeurs manquantes ont pu être regroupés en 8 classes présentées dans la figure B.2b.



(a) Base de données catégorielles

(b) Résultat de classification

FIGURE B.2 – Démarche de classification dans le cas des séquences comprenant des états manquants : (a) séquences d'habitudes de consommation issues de 1 778 compteurs et (b) 8 classes obtenues (séparée par les lignes rouges horizontales) en utilisant l'extension de la méthode proposée

Annexe C

Accélération de l'algorithme de classification FReMix

De manière générale, le temps d'exécution des algorithmes classiques de partitionnement de données tels que l'algorithme des k-means ou encore l'algorithme EM est connu pour être élevé quand le nombre d'unités à classifier s'accroît. Ce phénomène est encore plus prononcé pour des données de grande dimension comme c'est le cas notamment dans notre situation où chaque unité est une courbe constituée de plusieurs valeurs de consommation. Par exemple, l'utilisation de l'algorithme EM-FReMix pour classifier 500 000 courbes hebdomadaires (constituées chacune de 168 consommations) en 8 classes dure environ 36 heures sur un PC avec un processeur Intel Xeon (3,6Ghz CPU) et 64Go de RAM.

Ainsi, malgré leur finesse, les algorithmes de classification préalablement développés ne permettaient d'analyser qu'un échantillon de données de taille limitée. Pour faire face à cette difficulté, trois solutions sont possibles pour accélérer les algorithmes de classification :

- Réduire les données avant la classification en utilisant des méthodes de quantification adaptées, sans trop altérer la géométrie des données (ex. k-means lancé par exemple avec plusieurs centaines de classes) ;
- Effectuer la classification de manière incrémentale sans reconsidérer la masse de données déjà traitées ; on parle dans ce cas de méthode de classification en ligne ;
- Utiliser les technologies et architectures « Big Data » pour paralléliser et distribuer les calculs.

Dans cette thèse, nous avons opté pour la première et la troisième approche. Concernant la solution portant sur les plateformes dédiées aux calculs distribués, nous nous sommes basés sur les processeurs graphiques GPU qui sont de plus en plus utilisés pour la parallélisation des calculs scientifiques, et sur leur architecture de programmation CUDA.

Dans la suite, on décrit d'abord la première approche qui n'est qu'une version pondérée de l'algorithme FReMix.

C.1 Algorithme FReMix pondéré

Pour identifier les principaux profils saisonniers $((\mathbf{x}_{is})_{1\leq i\leq n;1\leq s\leq \mathrm{S}})$ issus d'un ensemble de 2 000 compteur portant sur une durée de 19 mois ($n \times \mathrm{S} = 1\,092\,000$ séries journalières à classifier), le temps de calcul de l'algorithme FReMix devient très élevé. Pour résoudre ce problème, nous proposons une approche qui consiste dans un premier temps à discrétiser les $n \times \mathrm{S}$ séries en 1 000 séries pondérées par l'algorithme de *k-means* (centres des classes obtenus avec $\mathrm{K} = 1\,000$), puis classifier ces nouvelles séries par une extension de l'algorithme FReMix qui prend en compte des pondérations dans les données. La réduction préalable des données via l'algorithme de *k-means*

Algorithme 5 : Algorithme EM pour l'extension du modèle FReMix (EM-FReMix pondéré)

Entrées : $n \times \mathrm{S}$ séries $(\mathbf{x}_{1s},\ldots,\mathbf{x}_{ns})_{1\leq s\leq \mathrm{S}}$, pondérations $(\kappa_1,\ldots,\kappa_n)$, nombre de composantes G,
paramètre initial $\theta^{(0)}$
$q \leftarrow 0$;

répéter

Étape E : calcul des probabilités a posteriori :

$$\tau_{ig}^{(q)} = \frac{p_g^{(q)}\mathcal{N}(\mathbf{x}_{is};\mathbf{U}\boldsymbol{\alpha}_g^{(q)},\sigma_g^{2^{(q)}}\mathbf{I})}{\sum_{h=1}^{\mathrm{G}} p_h^{(q)}\mathcal{N}(\mathbf{x}_{is};\mathbf{U}\boldsymbol{\alpha}_h^{(q)},\sigma_h^{2^{(q)}}\mathbf{I})}$$

Étape M : mise à jour des paramètres :

$$\begin{aligned}
p_k^{(q+1)} &\leftarrow \left(1/(\textstyle\sum_{i=1}^n \kappa_i)\right)\sum_{i=1}^n \kappa_i\tau_{ik}^{(q)} \\
\boldsymbol{\alpha}_k^{(q+1)} &\leftarrow \left[\left(\textstyle\sum_{i=1}^n \kappa_i\tau_{ik}^{(q)}\right)\sum_{t=1}^m \mathbf{u}_t\mathbf{u}_t'\right]^{-1}\left[\sum_{t=1}^m\left(\sum_{i=1}^n \kappa_i\tau_{ik}^{(q)}x_{it}\right)\mathbf{u}_t\right] \\
\sigma_k^{2^{(q+1)}} &\leftarrow \textstyle\sum_{i=1}^n \kappa_i\tau_{ik}^{(q)}\sum_{t=1}^m (x_{it}-\mathbf{u}_t'\boldsymbol{\alpha}_k^{(q+1)})^2/\left(m\sum_{i=1}^n \kappa_i\tau_{ik}^{(q)}\right)
\end{aligned}$$

$q \leftarrow q+1$

jusqu'à ce que la vraisemblance converge;

Sorties : paramètre $\hat{\theta}$

se comporte comme une méthode de quantification vectorielle. Le pseudo-code 5 détaille la procédure d'estimation des paramètres de l'extension de l'algorithme FReMix proposée.

Une fois que les 1 000 profils pondérés issus de l'algorithme de *k-means* sont regroupés à l'aide de l'algorithme proposé, on peut affecter a posteriori l'ensemble des $n \times \mathrm{S}$ séries initiales à des groupes obtenus.

C.2 Accélération à l'aide des unités de traitement graphique (GPU)

Effectuer des calculs via les cartes graphiques GPU (Graphical Processing Unit) permet d'accroître les performances des algorithmes en termes de temps d'exécution. Cela est dû principalement à la parallélisation des calculs. En utilisant cette technologie, les portions de code les plus lourdes en ressources de calcul sont gérées par le GPU, le reste des calculs étant affectés au CPU. Le mode de calcul par GPU, qui est de plus en plus répandu, nécessite de disposer d'un matériel adapté.

C.2.1 Généralités (architecture par bloc)

Contrairement aux CPU, les processeurs graphiques GPU sont construits de manière à ce que davantage de transistors soient dédiés au traitement des données plutôt qu'à la mémoire cache et au contrôle des flux, comme l'illustre la figure C.1.

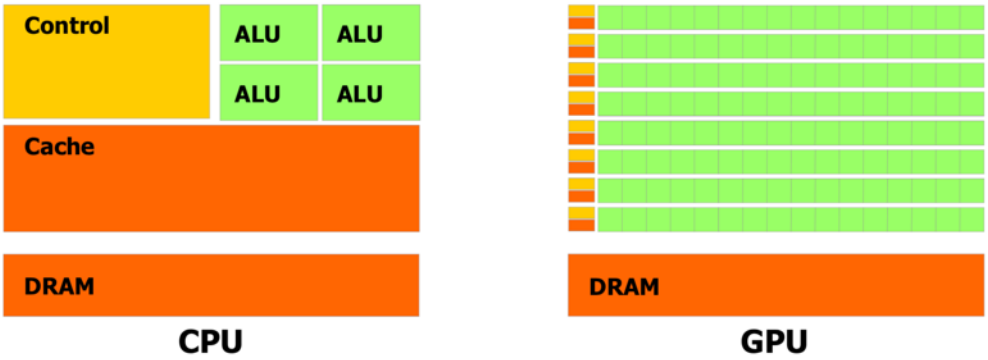


FIGURE C.1 – Architectures GPU et CPU : plus d'unités de calcul arithmétique (ALU) dans le cas du GPU

La différence fondamentale entre le CPU et le GPU se situe surtout dans leur manière de traiter chaque opération : les CPU incluent un nombre restreint de cœurs optimisés pour le traitement en série, alors que les GPU intègrent des milliers de cœurs conçus pour traiter efficacement de nombreuses tâches de manière simultanée.

Pour pouvoir exécuter des applications sur le GPU, le modèle de programmation parallèle CUDA (CUDA, 2012; NVIDIA, 2011) est le plus utilisé. Il permet de partitionner les données en plusieurs blocs (thread blocks), où chaque bloc est traité indépendamment et en parallèle. Chacun des blocs de données peut également être partitionné en pièces plus fines qui sont traitées en parallèle selon les threads existants. Les blocs sont organisés dans des grilles à une, deux ou trois dimensions. La figure C.2 montre une grille à deux dimensions.

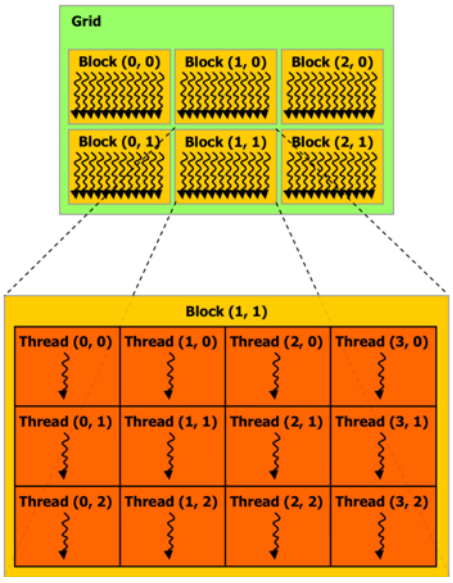


FIGURE C.2 – Grilles et blocs d'un GPU

Le nombre de blocs dans une grille dépend généralement de la taille des données à traiter et du nombre de processeurs du GPU. Les blocs d'une grille peuvent être planifiés sur tous les multiprocesseurs disponibles du GPU. Considérons, par exemple, le cas de données partitionnées en 8 blocs comme le montre la figure C.3. Selon le nombre de multiprocesseurs disponibles sur la carte GPU, les blocs sont doncpartagés automatiquement entre les multiprocesseurs existants afin d'être traités simultanément.

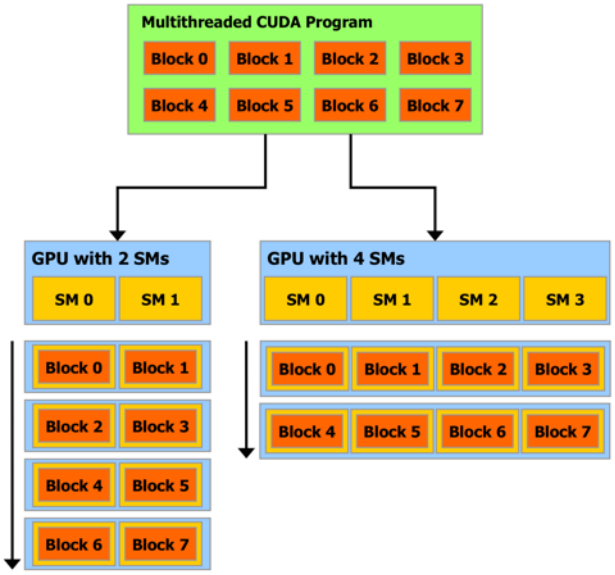


FIGURE C.3 – Partage automatique des blocs suivant les SM (Streaming Multiprocessors)

C.2.2 Hiérarchie de la mémoire GPU

Il existe plusieurs types de mémoire sur une carte graphique, comme le montre la figure C.4. Les plus importants sont :

- La mémoire constante, qui est celle où les variables constantes et les arguments du kernel sont enregistrés. Elle est assez lente mais elle dispose d'un cache de 8 kb ;
- La mémoire globale, qui est accessible par tous les blocs et leurs threads. Son temps de lecture et d'écriture pouvant être assez long, le temps gagné par l'accélération des calculs peut être perdu si l'on effectue un nombre élevé d'opérations sur cette mémoire ;
- La mémoire partagée, dont le temps de latence est très faible (jusqu'à 100 fois moindre que le temps de latence de la mémoire globale). Elle est accessible par tous les threads ;
- Les registres, qui constituent les mémoires les plus rapides mais qui sont dédiés à chaque thread ;
- La mémoire locale, qui est plus lente que la mémoire partagée et qui est utilisée pour toutes les opérations qui dépassent le cadre des registres.

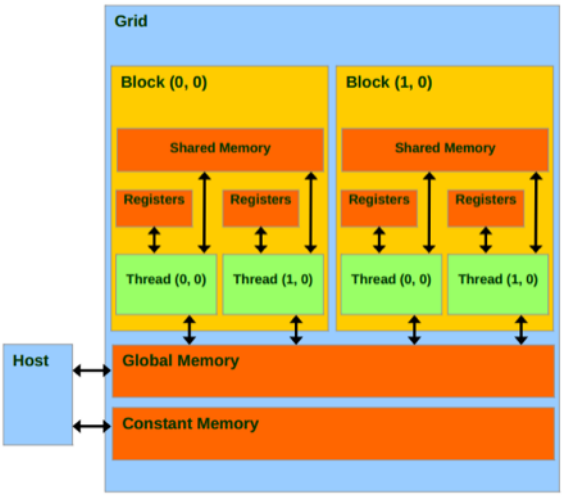


FIGURE C.4 – Hiérarchie de la mémoire GPU

C.2.3 Accélération de l'algorithme de clustering FReMix

Une analyse détaillée des opérations requises par l'algorithme EM-FReMix et de leur temps de calcul a mis en évidence que les multiplications de matrices étaient les opérations qui prenaient le plus de temps. Nous nous sommes donc focalisés précisément sur ces calculs. Sous le logiciel Matlab, plusieurs stratégies peuvent être employées pour accélérer les calculs matriciels en exploitant le GPU :

- les gpuArray,
- la programmation avec CUDA (Compute Unified Device Architecture).

Ces deux approches sont décrites par la suite.

gpuArray

L'approche gpuArray permet d'enregistrer les matrices de données directement dans la mémoire locale du GPU. Une fois le transfert des matrices effectué sur cette mémoire, des calculs usuels peuvent y être effectués ; l'optimisation de ces calculs est gérée automatiquement par Matlab.

Programmation avec CUDA

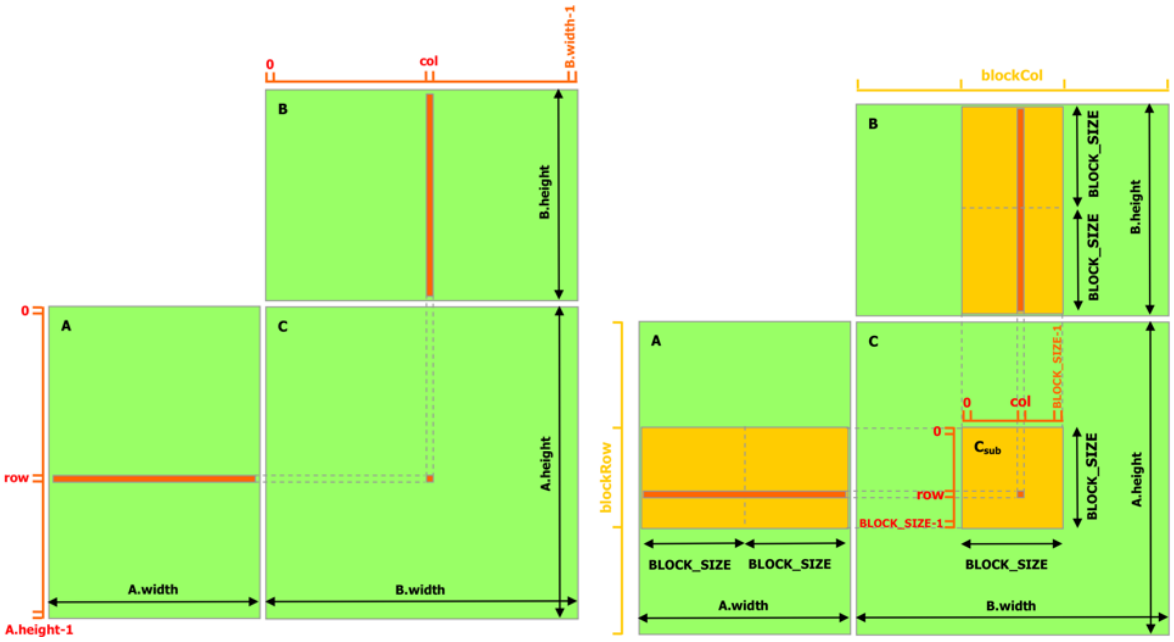
En utilisant le modèle de programmation CUDA (CUDA, 2012), le temps de calcul des opérations nécessitant d'importantes ressources telles le produit de grandes matrices peut être davantage optimisé. Il s'agit d'un code de bas niveau réalisé en langage C ou en Fortran qui permet une gestion plus fine de la mémoire. Notamment, il permet d'utiliser plus de mémoire partagée que de mémoire globale. L'utilisation de ce mode de programmation sous Matlab nécessite de coder en C une fonction appelée Kernel qui est ensuite compilée avec le compilateur NVIDIA dénommé NVCC afin de pouvoir être utilisée sous Matlab.

Focalisons-nous maintenant sur la multiplication matricielle qui est un point central dans l'accélération de l'algorithme FReMix. Les deux approches généralement utilisées dans la programmation CUDA, et que nous avons mises en œuvre pour l'algorithme EM-FReMix sont :

- la multiplication matricielle utilisant la mémoire globale. Cette approche, qui sera appelée CUDA1 dans la suite, utilise une implémentation classique de la multiplication matricielle. Chaque thread, comme le montre la figure C.5a, lit une ligne de la matrice A et une colonne de la matrice B et calcule l'élément correspondant dans la matrice C.
- la multiplication matricielle utilisant la mémoire partagée. Cette approche, appelé CUDA2 dans la suite, généralise le principe de calcul matriciel précédent en considérant des blocs de lignes et des blocs de colonnes (voir figure C.5b). Le produit d'un bloc ligne par un bloc colonne est parallélisé comme dans l'approche CUDA1.

C.2.4 Prise en compte de la mémoire limitée des cartes GPU

Bien que les cartes GPU permettent d'accélérer les calculs scientifiques, leur capacité en terme de mémoire est limitée. En outre, le prix d'une carte GPU augmente considérablement selon sa mémoire. Dans les travaux présentés dans ce rapport, nous avons utilisé la carte Nvidia Quadro M4000 qui dispose de 8 Go de mémoire. Avec cette taille de mémoire, nous avons pu réaliser la classification de 23 400 courbes hebdomadaires de consommation avec l'algorithme EM-FReMix. Cette capacité peut être augmentée si les matrices utilisées dans le programme sont découpées en blocs de dimension plus petite. Pour pouvoir traiter plus de données, nous avons combiné le CPU



(a) Mémoire globale (b) Mémoire partagée

FIGURE C.5 – Multiplication matricielle en utilisant la mémoire globale (a) et la mémoire partagée (b) de la carte graphique

et le GPU (combinaison qui sera appelée CPU+GPU par la suite). Dans cette approche, seuls les produits de grandes matrices sont effectués via le GPU, les autres opérations étant exécutées en utilisant le CPU. Un nombre maximum de 39 000 courbes hebdomadaires de consommation a pu ainsi être traité.

C.3 Évaluation des méthodes en termes de temps d'exécution

Les performances des méthodes suivantes ont été comparées :

- CPU(standard) : version standard (non accélérée) de l'algorithme EM-FReMix lancée via le CPU ;
- CPU(online) : version en ligne de l'algorithme EM-FReMix lancée via le CPU ;
- GPU(gpuarray) : version GPU de de l'algorithme EM-FReMix où le calcul matriciel est réalisé suivant l'approche gpuArray ;
- GPU(CUDA1) : version GPU de de l'algorithme EM-FReMix où le produit matriciel est réalisé suivant l'approche CUDA1 ;
- GPU(CUDA2) : version GPU de de l'algorithme EM-FReMix où le produit matriciel est réalisé suivant l'approche CUDA2 ;
- GPU+CPU(CUDA2) : version GPU+CPU de l'algorithme EM-FReMix où seuls les produits de grandes matrices ont été réalisés suivant l'approche CUDA2.

Pour effectuer ces comparaisons, la classification a été réalisée en considérant différentes tailles de données et deux nombres de classes : 8 et 16. Rappelons que les calculs ont été effectués sur un PC muni d'un processeur Intel Xeon (CPU 3,6GHz), de 64Go de RAM, et d'une carte graphique Nvidia M4000 avec 8 Go de mémoire et 1664 cœurs. Les résultats obtenus en termes de temps de calcul sont présentés dans le tableau C.1 et les figures C.6 et C.7.

En regardant le tableau C.1 et les graphiques C.6 et C.7 on peut voir que les méthodes online et GPU(CUDA2) sont globalement celles qui donnent les meilleurs temps de calcul. Elles sont à peu près 10 fois plus rapides que la méthode CPU(standard) et 5 fois plus rapides que la

TABLEAU C.1 – Temps de calcul (en minutes) de l'algorithme EM-FReMix et de ses différentes stratégies d'accélération, pour différentes tailles de courbes de consommation hebdomadaire; les NaN dans le tableau indiquent que les algorithmes sont dans l'incapacité de gérer la taille de données correspondante

Processeur	Méthode	7 800 courbes		15 600 courbes		23 400 courbes		31 200 courbes	
		K = 8	K = 16	K = 8	K = 16	K = 8	K = 16	K = 8	K = 16
CPU	standard	33,78	83,03	95,28	194,31	132,26	257,76	137,4	405,2
	online	4,30	**8,56**	**8,50**	17,55	13,24	**26,39**	**18,12**	35,50
GPU	gpuarray	17,63	21,71	26,58	28,26	43,4	NaN	NaN	NaN
	CUDA 1	6,65	12,45	10,86	20,56	14,53	NaN	NaN	NaN
	CUDA 2	**4,08**	10	9,98	**17,25**	**11,6**	NaN	NaN	NaN
CPU + GPU	CUDA 2	20,53	34,23	43,45	92,85	60,01	96,63	101,7	227,11

méthode GPU(gpuarray). La méthode GPU(CUDA2) est 2 fois plus rapide que la version optimisée GPU(CUDA1). Le seul inconvénient de cette méthode est qu'elle nécessite une taille de mémoire assez élevée de la carte GPU afin de pouvoir y charger des matrices de grande taille. L'approche CPU+GPU permet de traiter plus de données mais est moins performante que les autres approches exploitant le GPU. Elle est environ deux fois plus rapide que l'approche CPU.

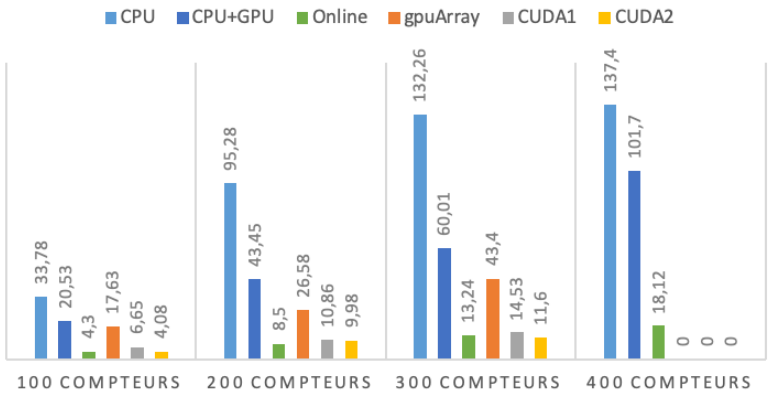


FIGURE C.6 – Temps de calcul (en minutes) de l'algorithme EM-FReMix et de ses différentes stratégies d'accélération en utilisant K = 8 classes

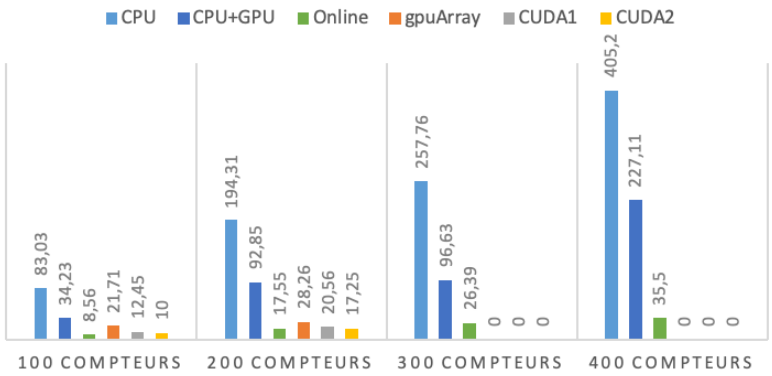


FIGURE C.7 – Temps de calcul (en minutes) de l'algorithme EM-FReMix et de ses différentes stratégies d'accélération en utilisant K = 16 classes

Bibliographie

ADAMOWSKI, J. F. 2008, «Peak daily water demand forecast modeling using artificial neural networks», *Journal of Water Resources Planning and Management*, vol. 134, nº 2, p. 119–128. 36

AGRESTI, A. et M. KATERI. 2011, *Categorical data analysis*, Springer. 33

AHMAD, S., A. LAVIN, S. PURDY et Z. AGHA. 2017, «Unsupervised real-time anomaly detection for streaming data», *Neurocomputing*, vol. 262, p. 134–147. 72

AKAIKE, H. 1974, «A new look at the statistical model identification», dans *Selected Papers of Hirotugu Akaike*, Springer, p. 215–222. 17

AKSELA, K. et M. AKSELA. 2010, «Demand estimation with automated meter reading in a distribution network», *Journal of Water Resources Planning and Management*, vol. 137, nº 5, p. 456–467. 34

ALTUNKAYNAK, A. et T. A. NIGUSSIE. 2017, «Monthly water consumption prediction using season algorithm and wavelet transform–based models», *Journal of Water Resources Planning and Management*, vol. 143, nº 6, p. 04017 011. 35

AMBROISE, C. et G. GOVAERT. 2000, «Clustering by maximizing a fuzzy classification maximum likelihood criterion», dans *COMPSTAT*, Springer, p. 187–192. 17

ARABIE, P. et G. DE SOETE. 1996, *Clustering and classification*, World Scientific. 14

AUE, A. et L. HORVÁTH. 2013, «Structural breaks in time series», *Journal of Time Series Analysis*, vol. 34, nº 1, p. 1–16. 8

BASSEVILLE, M. 1988, «Detecting changes in signals and systems—a survey», *Automatica*, vol. 24, nº 3, p. 309–326. 68

BASSEVILLE, M., I. V. NIKIFOROV et collab.. 1993, *Detection of abrupt changes : theory and application*, vol. 104, Prentice Hall Englewood Cliffs. 68, 70

BENGIO, Y. 1999, «Markovian models for sequential data», *Neural computing surveys*, vol. 2, nº 199, p. 129–162. 36

BENGIO, Y. et P. FRASCONI. 1996, «Input-output hmms for sequence processing», *IEEE Transactions on Neural Networks*, vol. 7, nº 5, p. 1231–1249. 36

BIERNACKI, C., G. CELEUX et G. GOVAERT. 2000, «Assessing a mixture model for clustering with the integrated completed likelihood», *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, nº 7, p. 719–725. 17, 40

BIERNACKI, C., G. CELEUX et G. GOVAERT. 2003, «Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models», *Computational Statistics & Data Analysis*, vol. 41, nº 3-4, p. 561–575. 16

BOUGADIS, J., K. ADAMOWSKI et R. DIDUCH. 2005, «Short-term municipal water demand forecasting», *Hydrological Processes : An International Journal*, vol. 19, n° 1, p. 137–148. 36

BOZDOGAN, H. 1987, «Model selection and akaike's information criterion (aic) : The general theory and its analytical extensions», *Psychometrika*, vol. 52, n° 3, p. 345–370. 17

BRITTON, T., G. COLE, R. STEWART et D. WISKAR. 2008, «Remote diagnosis of leakage in residential households», *Journal of Australian Water Association*, vol. 35, n° 6, p. 89–93. 7

CANDELIERI, A. et F. ARCHETTI. 2014, «Identifying typical urban water demand patterns for a reliable short-term forecasting–the icewater project approach», *Procedia Engineering*, vol. 89, p. 1004–1012. 36

CAO, Y., Y. LI, S. COLEMAN, A. BELATRECHE et T. M. MCGINNITY. 2015, «Adaptive hidden markov model with anomaly states for price manipulation detection», *IEEE transactions on neural networks and learning systems*, vol. 26, n° 2, p. 318–330. 80

CARDELL-OLIVER, R. 2013, «Water use signature patterns for analyzing household consumption using medium resolution meter data», *Water Resources Research*, vol. 49, n° 12, p. 8589–8599. 18

CARON, F., A. DOUCET et R. GOTTARDO. 2012, «On-line changepoint detection and parameter estimation with application to genomic data», *Statistics and Computing*, vol. 22, n° 2, p. 579–595. 72

CAVANAGH, S. M., W. M. HANEMANN et R. N. STAVINS. 2002, «Muffled price signals : household water demand under increasing-block prices», , n° 40. URL https://ssrn.com/abstract=317924. 18

CELEUX, G. et G. GOVAERT. 1992, «A classification em algorithm for clustering and two stochastic versions», *Computational statistics & Data analysis*, vol. 14, n° 3, p. 315–332. 15, 16, 38

CELEUX, G. et G. SOROMENHO. 1996, «An entropy criterion for assessing the number of clusters in a mixture model», *Journal of classification*, vol. 13, n° 2, p. 195–212. 17

CHEIFETZ, N. 2013, *Détection et classification de signatures temporelles CAN pour l'aide à la maintenance de sous-systèmes d'un véhicule de transport collectif*, thèse de doctorat, Paris Est. 77

CHEIFETZ, N., Z. NOUMIR, A. SAMÉ, A.-C. SANDRAZ, C. FÉLIERS et V. HEIM. 2017, «Modeling and clustering water demand patterns from real-world smart meter data», *Drinking Water Engineering and Science*, vol. 10, n° 2, doi :10.5194/dwes-10-75-2017, p. 75–82. URL https://www.drink-water-eng-sci.net/10/75/2017/. 21

CHEIFETZ, N., A. SAME, P. AKNIN, E. DE VERDALLE et D. CHENU. 2013, «A sequential testing procedure for multiple change-point detection in a stream of pneumatic door signatures», dans *2013 12th International Conference on Machine Learning and Applications*, vol. 1, IEEE, p. 117–122. 72

CHEN, S., P. GOPALAKRISHNAN et collab.. 1998, «Speaker, environment and channel change detection and clustering via the bayesian information criterion», dans *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8, Virginia, USA, p. 127–132. 71, 72

CHEN, X. C., K. STEINHAEUSER, S. BORIAH, S. CHATTERJEE et V. KUMAR. 2013, «Contextual time series change detection», dans *Proceedings of the 2013 SIAM International Conference on Data Mining*, SIAM, p. 503–511. 8

CHO, H. et collab.. 2016, «Change-point detection in panel data via double cusum statistic», *Electronic Journal of Statistics*, vol. 10, n° 2, p. 2000–2038. 77

CHO, S.-J., A. S. COHEN et B. BOTTGE. 2013, «Detecting intervention effects using a multilevel latent transition analysis with a mixture irt model», *Psychometrika*, vol. 78, n° 3, p. 576–600. 33

COLE, G. et R. A. STEWART. 2013, «Smart meter enabled disaggregation of urban peak water demand : precursor to effective urban water planning», *Urban Water Journal*, vol. 10, n° 3, p. 174–194. 5

COLLINS, L. M. et S. T. LANZA. 2009, *Latent class and latent transition analysis : With applications in the social, behavioral, and health sciences*, vol. 718, John Wiley & Sons. 96

CUDA, C. 2012, «Best practices guide», *Nvidia Corporation*. 107, 109

CUTLER, A. et M. P. WINDHAM. 1994, «Information-based validity functionals for mixture analysis», dans *Proceedings of the first US/Japan Conference on the Frontiers of statistical modeling : An informational approach*, Springer, p. 149–170. 17

DAY, H. et K. CONWAY. 2009, «Rule 1 : No watts no water, rule 2 : No water no watts», dans *Proceedings of 24th Annual WasteResue Symposium : Where Has All the Water Gone*. 4

DEMPSTER, A. P., N. M. LAIRD et D. B. RUBIN. 1977, «Maximum likelihood from incomplete data via the em algorithm», *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 39, n° 1, p. 1–22. 15, 38

DOMENE, E. et D. SAURÍ. 2006, «Urbanisation and water consumption : Influencing factors in the metropolitan region of barcelona», *Urban Studies*, vol. 43, n° 9, p. 1605–1623. 30

EDWARDS, R. E., J. NEW et L. E. PARKER. 2012, «Predicting future hourly residential electrical consumption : A machine learning case study», *Energy and Buildings*, vol. 49, p. 591–603. 35

ESTER, M., H.-P. KRIEGEL, J. SANDER, X. XU et collab.. 1996, «A density-based algorithm for discovering clusters in large spatial databases with noise.», dans *Kdd*, vol. 96, p. 226–231. 14

EVERETT, B. 2013, *An introduction to latent variable models*, Springer Science & Business Media. 32

FEARNHEAD, P. et Z. LIU. 2007, «On-line inference for multiple changepoint problems», *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 69, n° 4, p. 589–605. 72

FIELDING, K. S., A. SPINKS, S. RUSSELL, R. MCCREA, R. STEWART et J. GARDNER. 2013, «An experimental test of voluntary strategies to promote urban water demand management», *Journal of environmental management*, vol. 114, p. 343–351. 5

FLATH, C., D. NICOLAY, T. CONTE, C. VAN DINTHER et L. FILIPOVA-NEUMANN. 2012, «Cluster analysis of smart metering data», *Business & Information Systems Engineering*, vol. 4, n° 1, p. 31–39. 8

FOKIANOS, K., E. GOMBAY et A. HUSSEIN. 2014, «Retrospective change detection for binary time series models», *Journal of Statistical Planning and Inference*, vol. 145, p. 102–112. 73

FOKIANOS, K. et B. KEDEM. 1998, «Prediction and classification of non-stationary categorical time series», *Journal of multivariate analysis*, vol. 67, n° 2, p. 277–296. 35

FRALEY, C. et A. E. RAFTERY. 2002, «Model-based clustering, discriminant analysis, and density estimation», *Journal of the American statistical Association*, vol. 97, n° 458, p. 611–631. 14

FROEHLICH, J., E. LARSON, S. GUPTA, G. COHN, M. REYNOLDS et S. PATEL. 2010, «Disaggregated end-use energy sensing for the smart grid», *IEEE Pervasive Computing*, vol. 10, n° 1, p. 28–39. 5

GAFFNEY, S. et P. SMYTH. 1999, «Trajectory clustering with mixtures of regression models», dans *KDD*, vol. 99, p. 63–72. 21

GAGLIARDI, F., S. ALVISI, Z. KAPELAN et M. FRANCHINI. 2017, «A probabilistic short-term water demand forecasting model based on the markov chain», *Water*, vol. 9, nº 7, p. 507. 36

GAUVAIN, J.-L. et C.-H. LEE. 1994, «Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains», *IEEE transactions on speech and audio processing*, vol. 2, nº 2, p. 291–298. 39

GAY, D., R. GUIGOURÈS, M. BOULLÉ et F. CLÉROT. 2015, «Cats & co : Categorical time series coclustering», *arXiv preprint arXiv :1505.01300*. 96

GOMBAY, E., F. LI et H. YU. 2017, «Retrospective change detection in categorical time series», *Communications in Statistics-Theory and Methods*, vol. 46, nº 14, p. 6831–6845. 73, 80

GOVAERT, G. et M. NADIF. 2003, «Clustering with block mixture models», *Pattern Recognition*, vol. 36, nº 2, p. 463–473. 96

GREEN, P. J. 1984, «Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives», *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 46, nº 2, p. 149–170. 76

GUSTAFSSON, F. 2000, *Adaptive filtering and change detection*, vol. 1, Citeseer. 68

GUTZLER, D. S. et J. S. NIMS. 2005, «Interannual variability of water demand and summer climate in albuquerque, new mexico», *Journal of Applied Meteorology*, vol. 44, nº 12, p. 1777–1787. 18

HABEN, S., C. SINGLETON et P. GRINDROD. 2015, «Analysis and clustering of residential customers energy behavioral demand using smart meter data», *IEEE transactions on smart grid*, vol. 7, nº 1, p. 136–144. 8

HANKE, S. H. et L. D. MARE. 1982, «Residential water demand : A pooled, time series, cross section study of malmö, sweden 1», *JAWRA Journal of the American Water Resources Association*, vol. 18, nº 4, p. 621–626. 18

HARRINGTON, P. 2012, *Machine learning in action*, Manning Publications Co. 8

HÖHLE, M. 2010, «Online change-point detection in categorical time series», dans *Statistical modelling and regression structures*, Springer, p. 377–397. 73

HOLLAND, P. W. et R. E. WELSCH. 1977, «Robust regression using iteratively reweighted least-squares», *Communications in Statistics-theory and Methods*, vol. 6, nº 9, p. 813–827. 35, 39, 99

HOUSE-PETERS, L., B. PRATT et H. 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*, vol. 46, nº 3, p. 461–472. 30

HOUSE-PETERS, L. A. et H. CHANG. 2011, «Urban water demand modeling : Review of concepts, methods, and organizing principles», *Water Resources Research*, vol. 47, nº 5. 7

HUBERT, L. et P. ARABIE. 1985, «Comparing partitions», *Journal of classification*, vol. 2, nº 1, p. 193–218. 43

HUNAIDI, O., A. WANG, M. BRACKEN, T. GAMBINO et C. FRICKE. 2005, «Detecting leaks in water distribution pipes», *Arab Water World*, vol. 29, nº 4, p. 52–55. 4

ISERMANN, R. 1984, «Process fault detection based on modeling and estimation methods—a survey», *automatica*, vol. 20, nº 4, p. 387–404. 68

JAIN, A., A. K. VARSHNEY et U. C. JOSHI. 2001, «Short-term water demand forecast modelling at iit kanpur using artificial neural networks», *Water resources management*, vol. 15, nº 5, p. 299–321. 35

JONES, C. V. et J. R. MORRIS. 1984, «Instrumental price estimates and residential water demand», *Water Resources Research*, vol. 20, nº 2, p. 197–202. 18

JUNG, J., V. PAXSON, A. W. BERGER et H. BALAKRISHNAN. 2004, «Fast portscan detection using sequential hypothesis testing», dans *IEEE Symposium on Security and Privacy, 2004. Proceedings. 2004*, IEEE, p. 211–225. 74

KASS, R. E. et A. E. RAFTERY. 1995, «Bayes factors», *Journal of the american statistical association*, vol. 90, nº 430, p. 773–795. 17

KENNEY, D. S., C. GOEMANS, R. KLEIN, J. LOWREY et K. REIDY. 2008, «Residential water demand management : lessons from aurora, colorado 1», *JAWRA Journal of the American Water Resources Association*, vol. 44, nº 1, p. 192–207. 8

KEOGH, E., K. CHAKRABARTI, M. PAZZANI et S. MEHROTRA. 2001, «Dimensionality reduction for fast similarity search in large time series databases», *Knowledge and information Systems*, vol. 3, nº 3, p. 263–286. 8

KUMAR, U. et V. JAIN. 2010, «Time series models (grey-markov, grey model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in india», *Energy*, vol. 35, nº 4, p. 1709–1716. 35

KWAC, J., J. FLORA et R. RAJAGOPAL. 2014, «Household energy consumption segmentation using hourly data», *IEEE Transactions on Smart Grid*, vol. 5, nº 1, p. 420–430. 33

LABEEUW, W. et G. DECONINCK. 2013, «Residential electrical load model based on mixture model clustering and markov models», *IEEE Transactions on Industrial Informatics*, vol. 9, nº 3, p. 1561–1569. 33

LANZA, S. T. et L. M. COLLINS. 2008, «A new sas procedure for latent transition analysis : Transitions in dating and sexual risk behavior.», *Developmental psychology*, vol. 44, nº 2, p. 446. 33

LAZARSFELD, P. F. et N. W. HENRY. 1968, *Latent structure analysis*, Houghton Mifflin Co. 32

LI, F., A. COHEN, B. BOTTGE et J. TEMPLIN. 2016, «A latent transition analysis model for assessing change in cognitive skills», *Educational and Psychological Measurement*, vol. 76, nº 2, p. 181–204. 33

LI, J., F. TSUNG et C. ZOU. 2013, «Directional change-point detection for process control with multivariate categorical data», *Naval Research Logistics (NRL)*, vol. 60, nº 2, p. 160–173. 73

LIN, J., E. KEOGH, S. LONARDI et B. CHIU. 2003, «A symbolic representation of time series, with implications for streaming algorithms», dans *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, ACM, p. 2–11. 8, 34, 96

LORDEN, G. et collab.. 1971, «Procedures for reacting to a change in distribution», *The Annals of Mathematical Statistics*, vol. 42, nº 6, p. 1897–1908. 70, 71

LUONG, T. M., V. PERDUCA et G. NUEL. 2012, «Hidden markov model applications in change-point analysis», *arXiv preprint arXiv :1212.1778*. 73

MACQUEEN, J. et collab.. 1967, «Some methods for classification and analysis of multivariate observations», dans *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, Oakland, CA, USA, p. 281–297. 14

MARTIN, R. A., W. F. VELICER et J. L. FAVA. 1996, «Latent transition analysis to the stages of change for smoking cessation», *Addictive Behaviors*, vol. 21, nº 1, p. 67–80. 33

MAYER, P. W., W. B. DEOREO, E. M. OPITZ, J. C. KIEFER, W. Y. DAVIS, B. DZIEGIELEWSKI et J. O. NELSON. 1999, «Residential end uses of water», . 5, 7

MCCUTCHEON, A. L. 1987, *Latent class analysis*, 64, Sage. 32

MCINTYRE, T. 2008, «Data retention in ireland : Privacy, policy and proportionality», *Computer Law & Security Review*, vol. 24, nº 4, p. 326–334. 7

MCKENNA, S., F. FUSCO et B. ECK. 2014, «Water demand pattern classification from smart meter data», *Procedia Engineering*, vol. 70, p. 1121–1130. 34

MCLACHLAN, G. et T. KRISHNAN. 2007, *The EM algorithm and extensions*, vol. 382, John Wiley & Sons. 15

MCLACHLAN, G. J. et K. E. BASFORD. 1988, *Mixture models : Inference and applications to clustering*, vol. 84, M. Dekker New York. 14

MELZI, F., A. SAMÉ, M. ZAYANI et L. OUKHELLOU. 2017, «A dedicated mixture model for clustering smart meter data : identification and analysis of electricity consumption behaviors», *Energies*, vol. 10, nº 10, p. 1446. 33

MOUSTAKIDES, G. V. et collab.. 1986, «Optimal stopping times for detecting changes in distributions», *The Annals of Statistics*, vol. 14, nº 4, p. 1379–1387. 70

NADIF, M. et G. GOVAERT. 2005, «Block clustering of contingency table and mixture model», dans *International Symposium on Intelligent Data Analysis*, Springer, p. 249–259. 96

NVIDIA, C. 2011, «Nvidia cuda c programming guide», *Nvidia Corporation*, vol. 120, nº 18, p. 8. 107

ORACLE. 2009, «Smart metering for water utilities», . 7

PAGE, E. S. 1954, «Continuous inspection schemes», *Biometrika*, vol. 41, nº 1/2, p. 100–115. 70

RABINER, L. R. 1989, «A tutorial on hidden markov models and selected applications in speech recognition», *Proceedings of the IEEE*, vol. 77, nº 2, p. 257–286. 73

ROOS, K., T. TERLAKY et J.-P. VIAL. 1998, «Theory and algorithms for linear optimization - an interior point approach», dans *Wiley-Interscience series in discrete mathematics and optimization*. 39, 99

ROWLANDS, I. H., T. REID et P. PARKER. 2015, «Research with disaggregated electricity end-use data in households : review and recommendations», *Wiley Interdisciplinary Reviews : Energy and Environment*, vol. 4, nº 5, p. 383–396. 7

SAMÉ, A., C. AMBROISE et G. GOVAERT. 2007, «An online classification em algorithm based on the mixture model», *Statistics and Computing*, vol. 17, nº 3, p. 209–218. 17, 96

SAMÉ, A. et G. GOVAERT. 2012, «Online time series segmentation using temporal mixture models and bayesian model selection», dans *2012 11th International Conference on Machine Learning and Applications*, vol. 1, IEEE, p. 602–605. 72

SAMÉ, A., Z. NOUMIR, N. CHEIFETZ, A.-C. SANDRAZ et C. FÉLIERS. 2016, «Décomposition et classification de données fonctionnelles pour l'analyse de la consommation d'eau», . 21, 22

SATO, M.-A. et S. ISHII. 2000, «On-line em algorithm for the normalized gaussian network», *Neural computation*, vol. 12, n° 2, p. 407–432. 96

SCHWARZ, G. et collab.. 1978a, «Estimating the dimension of a model», *The annals of statistics*, vol. 6, n° 2, p. 461–464. 17

SCHWARZ, G. et collab.. 1978b, «Estimating the dimension of a model», *The annals of statistics*, vol. 6, n° 2, p. 461–464. 71

SIMONOFF, J. S. 2012, *Smoothing methods in statistics*, Springer Science & Business Media. 88

SMYTH, P. 1999, «Probabilistic model-based clustering of multivariate and sequential data», dans *Proceedings of the Seventh International Workshop on AI and Statistics*, San Francisco, CA : Morgan Kaufman, p. 299–304. 32

SOROMENHO, G. 1994, «Comparing approaches for testing the number of components in a finite mixture model», *Computational Statistics*, vol. 9, n° 1, p. 65–78. 17

STEWART, R. A., R. WILLIS, D. GIURCO, K. PANUWATWANICH et G. CAPATI. 2010, «Web-based knowledge management system : linking smart metering to the future of urban water planning», *Australian Planner*, vol. 47, n° 2, p. 66–74. 7

THOMAS, J. F. et G. J. SYME. 1988, «Estimating residential price elasticity of demand for water : A contingent valuation approach», *Water Resources Research*, vol. 24, n° 11, p. 1847–1857. 7

VELICER, W. F., R. A. MARTIN et L. M. COLLINS. 1996, «Latent transition analysis for longitudinal data», *Addiction*, vol. 91, n° 12s1, p. 197–210. 33

WALKER, D., E. CREACO, L. VAMVAKERIDOU-LYROUDIA, R. FARMANI, Z. KAPELAN et D. SAVIĆ. 2015, «Forecasting domestic water consumption from smart meter readings using statistical methods and artificial neural networks», *Procedia Engineering*, vol. 119, p. 1419–1428. 35

WANG, T., G.-L. LU, J. LIU et P. YAN. 2017, «Adaptive change detection for long-term machinery monitoring using incremental sliding-window», *Chinese Journal of Mechanical Engineering*, vol. 30, n° 6, p. 1338–1346. 78

WANG, Y., Q. CHEN, C. KANG et Q. XIA. 2016, «Clustering of electricity consumption behavior dynamics toward big data applications», *IEEE transactions on smart grid*, vol. 7, n° 5, p. 2437–2447. 33, 34

WARD JR, J. H. 1963, «Hierarchical grouping to optimize an objective function», *Journal of the American statistical association*, vol. 58, n° 301, p. 236–244. 14

WERBOS, P. J. 2011, «Computational intelligence for the smart grid-history, challenges, and opportunities», *IEEE Computational Intelligence Magazine*, vol. 6, n° 3, p. 14–21. 4

WILLIS, R. M., R. A. STEWART, K. PANUWATWANICH, P. R. WILLIAMS et A. L. HOLLINGSWORTH. 2011, «Quantifying the influence of environmental and water conservation attitudes on household end use water consumption», *Journal of environmental management*, vol. 92, n° 8, p. 1996–2009. 30

WILLSKY, A. et H. JONES. 1976, «A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems», *IEEE Transactions on Automatic control*, vol. 21, n° 1, p. 108–112. 77

ZHOU, S., T. MCMAHON, A. WALTON et J. LEWIS. 2002, «Forecasting operational demand for an urban water supply zone», *Journal of hydrology*, vol. 259, nº 1-4, p. 189–202. 30

ZHOU, S. L., T. A. MCMAHON, A. WALTON et J. LEWIS. 2000, «Forecasting daily urban water demand : a case study of melbourne», *Journal of hydrology*, vol. 236, nº 3-4, p. 153–164. 30

Liste des publications

Revues

- Milad Leyli-Abadi, Allou Samé, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cédric Féliers, et Olivier Chesneau. « Mixture of Joint Nonhomogeneous Markov Chains to Cluster and Model Water Consumption Behavior Sequences ». Dans *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2019.
- Milad Leyli-Abadi, Allou Samé, Latifa Oukhellou, Nicolas Cheifetz et Pierre Mandel. « Online change point detection for categorical time series using an adaptive threshold ». En soumission dans « Neurocomputing »

Conférences internationales

- Allou Samé, Milad Leyli-Abadi et Latifa Oukhellou. « Change detection in smart grids using dynamic mixtures of t-distributions ». Dans *World Congress on Condition Monitoring (WCCM)*, 2019.
- Allou Samé et Milad Leyli-Abadi. « Change Point Detection in Periodic Panel Data Using a Mixture-Model-Based Approach ». Dans *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, Cham, 2019.
- Milad Leyli-Abadi, Allou Samé, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cédric Féliers, et Olivier Chesneau. « Mixture of Non-homogeneous Hidden Markov Models for Clustering and Prediction of Water Consumption Time Series ». Dans *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8., 2018.
- Milad Leyli-Abadi, Allou Samé, Latifa Oukhellou, Nicolas Cheifetz, Pierre Mandel, Cédric Feliers, et Olivier Chesneau. « Predictive classification of water consumption time series using non-homogeneous markov models ». Dans *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017.
- Milad Leyli-Abadi, Lazhar Labiod, et Mohamed Nadif. « Denoising autoencoder as an effective dimensionality reduction and clustering of text data ». Dans *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, Cham, 2017.

Conférence nationale

- Milad Leyli-abadi, Allou Samé, and Latifa Oukhellou. « Détection en ligne de multiples changements dans un panel de données catégorielles ». Dans *XXVIèmes Rencontres de la Société Francophone de Classification*, SFC 2019.
- Milad Leyli-Abadi, Allou Samé et Latifa Oukhellou. « Mélange de chaînes de Markov non-homogènes pour la classification et la prévision des habitudes de consommation issues d'un réseau d'eau intelligent ». Dans *XXVèmes Rencontres de la Société Francophone de Classification*, SFC 2018.