Classification et décomposition de séries temporelles avec prise en compte de facteurs observables et inobservables : application à l'analyse des dynamiques de comportements d'occupants de logements à partir de données thermiques

Louise Bonfils

Classification et décomposition de séries temporelles avec prise en compte de facteurs observables et inobservables. Application à l'analyse des dynamiques de comportements d'occupants de logements à partir de données thermiques.

THÈSE DE DOCTORAT

Présentée et soutenue publiquement le 8 décembre 2022

pour l'obtention du grade de

Docteur de l'Université Gustave Eiffel

Ecole doctorale n° 532 - MSTIC
(Spécialité : Signal, Image, Automatique)

par

Louise Bonfils

Composition du jury :

NADIF Mohamed	Professeur à l'Université Paris Cité	*Président du jury*
DERRODE Stéphane	Professeur à l'École Centrale de Lyon	*Rapporteur*
OUDRE Laurent	Professeur à l'ENS Paris Saclay	*Rapporteur*
IONESCU Anda	Maître de Conférences à l'Université Paris-Est Créteil	*Examinatrice*
SAMÉ Allou	Directeur de recherche à l'Université Gustave Eiffel	*Directeur de thèse*
OUKHELLOU Latifa	Directrice de recherche à l'Université Gustave Eiffel	*Co-Directrice de thèse*
WAEYTENS Julien	Directeur de Recherche à l'Université Gustave Eiffel	*Invité*

Remerciements

Tout d'abord, je souhaite adresser mes plus sincères et profonds remerciements à mes deux directeurs de thèse Allou Samé et Latifa Oukhellou qui m'ont soutenue, accompagnée et conseillée tout au long de mes travaux. Vous avez été d'un immense soutien scientifique, d'une grande patience et toujours bienveillants durant ces trois années.

Tous mes remerciements et ma gratitude vont aussi à l'ensemble des membres du projet I-site Future ANDRE qui ont porté un regard extérieur et critique sur mes travaux. Ma reconnaissance va tout spécialement à Julien Waeytens qui m'a fait bénéficier de ses compétences dans le domaine de la thermique du bâtiment.

Je remercie également tous les membres du jury qui ont accepté d'évaluer mes travaux de thèse. Je parle de Stéphane Derrode et Laurent Oudre qui me font l'honneur d'être rapporteur de cette thèse, ainsi que de Mohamed Nadif, Anda Ionescu et à nouveau de Julien Waeytens.

Une pensée et un grand merci pour toutes les personnes que j'ai pu côtoyer au Grettia avec lesquelles j'ai partagé beaucoup de repas et de pauses-café au cours de ces trois dernières années.

Pour finir, je remercie mes parents pour m'avoir accompagnée tout au long de mes études. Et enfin, j'adresse également un grand merci à mes amis et à mon entourage pour leur présence à mes côtés depuis toutes ces années et pour la joie et le bonheur qu'ils me procurent jour après jour.

Résumé

Les travaux de cette thèse s'inscrivent dans le cadre de l'analyse de données thermiques et plus particulièrement de la classification de données pour l'extraction de comportement d'occupants au sein de bâtiments. Dans la mesure où ces comportements, constitués d'habitudes de chauffage, de consommation, ou encore de présence d'habitants au sein de leur logement, représentent une part importante de la consommation énergétique du secteur résidentiel, il est utile de bien les caractériser pour améliorer les prévisions de la consommation d'énergie dans ce secteur. Dans le cadre des travaux de cette thèse, on s'intéresse aux dynamiques de comportements dans le sens où ces derniers peuvent évoluer au cours du temps. Ces dynamiques comportementales dépendent de plusieurs facteurs et sont, par conséquent, variées et difficilement observables. L'objectif est donc de développer une méthode de classification non supervisée de données thermiques dans le but d'en extraire une typologie de comportements dynamiques d'occupants de bâtiments.

L'approche proposée dans ces travaux est basée sur un modèle probabiliste à variables latentes dynamiques. Son principal intérêt réside dans la séparation des effets relatifs à des facteurs contextuels d'une part et à la dynamique comportementale des occupants d'autre part. Nous proposons donc, dans ces travaux, un modèle de classification qui permet d'estimer des effets régressifs ainsi que des centres de classes, modélisés comme des processus stochastiques. Ce modèle, de par sa complexité, ne peut être estimé via des méthodes de maximum de vraisemblance habituelles. Par conséquent, on présentera un algorithme itératif basé sur des méthodes d'inférence variationnelle dans le but d'estimer les paramètres du modèle.

La méthode est évaluée sur deux ensembles de données. D'abord, les performances du modèle sont testées et comparées à des méthodes standard à partir d'un ensemble de données simulées via des modèles statistiques. Ensuite, l'application du modèle à un jeu de données issu d'un modèle thermique permet d'évaluer cette méthode dans un cadre d'analyse plus réaliste. Pour finir, la classification de données de température intérieure pour un ensemble de maisons via le modèle proposé permet de construire des classes homogènes. Les résultats permettent de mettre en évidence que les dynamiques de comportements estimées sont corrélées à des comportements de présence des habitants au sein de leur logement.

Mots clés : Classification non supervisée, Séries temporelles, Modèle dynamique à variables latentes, Approximation par inférence variationnelle, Données thermiques, Comportement des occupants dans les bâtiments.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Contents

1.1 Contexte applicatif

Avec l'importance des nouveaux enjeux climatiques, la question de la réduction de la consommation d'énergie est devenu un sujet majeur. La grande variété des sources de consommation d'énergie rend cette problématique complexe. En effet, la consommation d'énergie concerne les secteurs résidentiels, tertiaires, industriels ou encore celui des transports. De plus, au sein d'un même secteur, les sources de consommation sont également très nombreuses. Par exemple, dans le secteur résidentiel, le chauffage, l'utilisation d'appareil électroménager, la ventilation et encore l'éclairage sont des sources possibles de consommation d'énergie. Ce secteur représente, à lui seul, 36 % de la consommation finale d'électricité en France, en 2019 (1.1).

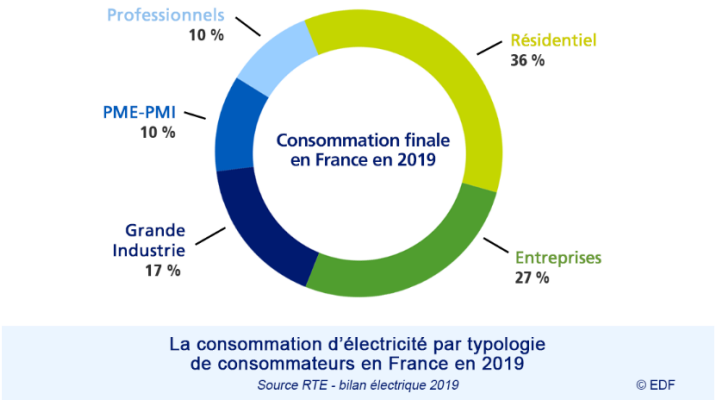

FIGURE 1.1 – Répartition de la consommation finale d'électricité en France en 2019, d'après le Bilan Énergétique RTE. Source : EDF

La réduction de la consommation d'énergie dans le secteur résidentiel passe généralement par la rénovation des bâtiments et la construction de nouveaux logements qui offrent de meilleures performances thermiques. Cet objectif peut également être atteint par une meilleure adéquation entre les logements, les appareils et les besoins des habitants. Par exemple, des systèmes de chauffage mieux adaptés aux besoins des habitants peuvent permettre d'éviter de chauffer des logements inoccupés. Il en est de même des systèmes de ventilation qui peuvent être adaptés pour ne fonctionner que lorsque cela est nécessaire. Ces évolutions nécessitent de meilleures connaissances des besoins et des comportements des habitants au sein de leur logement. La caractérisation de ces comportements est un sujet majeur de recherche tant les comportements sont variés et difficilement observables ou modélisables. En effet, comprendre et modéliser des comportements d'occupant au sein de logements peut permettre de construire des modèles de prédiction de consommation d'énergie plus performants. La modélisation des comportements peut également être utile dans le cadre de la simulation de la consommation d'un bâtiments afin de mieux prendre en compte les interactions des habitants avec les logements. Le concept de comportement est vaste et peut faire référence à différents éléments. En effet, on peut s'intéresser aux comportements de

consommation d'électricité, ou aux comportements de chauffage, aux comportements de présences au sein des logements ou encore aux activités des habitants au sein de leur habitat. Ces comportements, aussi variés soient-ils, ne sont pas toujours observables et mesurables. Par exemple, la présence des occupants au sein de leur logement n'est pas une variable facilement mesurable, ce qui rend l'analyse de ces comportements difficile. De plus, la variabilité et la large diversité des comportements constituent un autre obstacle à l'analyse de données de consommation d'énergie ou de données thermiques. Ainsi, si l'on souhaite comprendre, décrire et éventuellement modéliser certains comportements, il peut alors être intéressant de résumer ces comportements en un petit nombre de groupes représentatifs. Pour cela, la classification non supervisée de données est souvent utilisée car elle permet de construire des groupes homogènes de données afin de décrire des comportements type. Les travaux qui seront présentés par la suite s'inscrivent dans ce cadre. L'objectif est de proposer une méthode de classification de données thermique dans le but d'extraire des comportements d'habitants au sein de leur logement. Le choix est fait d'extraire des comportements à partir de données thermiques car ces données sont appropriées pour identifier des comportements d'occupation ou de chauffage. En effet, la température intérieure est une variable d'ambiance qui est impactée par le chauffage du logement, mais également par l'activité et l'occupation des habitants.

L'objectif de ces travaux est de développer une méthode de classification qui a pour intérêt de classifier des données temporelles, tout en identifiant des effets régressifs liés à des facteurs contextuels et en caractérisant les clusters par des processus stochastiques. La méthode qui est développée dans ces travaux s'inscrit dans le cadre des modèles de mélange à variables latentes. Son principal intérêt est qu'elle permet d'estimer des effets régressifs ainsi que des processus stochastiques caractérisant des clusters, à la différence de modèles de mélange de régression par exemple qui ne permettent d'identifier que des effets liés à des facteurs exogènes. Dans un cadre applicatif, l'objectif est d'identifier des dynamiques de comportements type à partir de données thermiques tout en identifiant les effets exogènes liés à des variables météorologiques ou calendaires. En effet, on s'intéresse plus particulièrement aux dynamiques de comportements dans le sens où ces derniers peuvent évoluer au cours du temps suite à des changements de contexte, de prix de l'énergie ou même d'habitants.

1.2 Les apports des travaux de thèse

Les apports des travaux de thèse présentés dans ce manuscrit :

- Tout d'abord, on construit un modèle de classification permettant de classifier des données temporelles de température mesurée au sein d'un ensemble de logements, d'estimer l'effet des facteurs exogènes communs et de modéliser les profils des clusters en tant que processus stochastiques. De plus, une variante de ce premier modèle est également proposée dans le cas où les effets régressifs sont propres à chaque cluster. Ce

modèle s'inscrit dans le cadre des modèles de mélange à variables latentes dynamiques et effets régressifs.

- Ensuite, les modèles proposés sont évalués et comparés à des modèles de référence construits à partir de méthodes standards. L'évaluation des performances se base sur trois critères relatifs aux erreurs d'estimation réalisées sur les centres de classe, sur l'effet des facteurs exogènes et sur la classification des observations.
- La motivation première étant la classification de données énergétiques et thermiques, les modèles sont appliqués à des jeux de données simulées à l'aide d'un modèle de simulation thermique du bâtiment. Cette étape permet de construire un jeu constitué de données réalistes, mais contrôlées.
- Enfin, la base de données en libre accès REFIT [Firth et al., 2017], constituée de données thermiques mesurées pour un ensemble de maisons individuelles permet d'identifier des classes de comportements type relatifs à des habitudes d'occupation des logements. La base de données ANDRE, regroupant des données de température mesurée au sein de parties communes de bâtiments permet d'identifier des classes de capteurs selon leur sensibilité aux conditions extérieures.

1.3 Organisation du manuscrit

Ce manuscrit est organisé en quatre chapitres.

Le premier chapitre est dédié à la présentation du contexte applicatif de ces travaux en passant en revue les méthodes d'analyse des données énergétiques et plus spécifiquement de la classification de données pour extraire des comportements d'habitants.

Le second chapitre permet de présenter les méthodes et concepts méthodologiques qui seront utilisés dans la suite des travaux. Il permet notamment d'introduire les modèles de mélange ainsi que les méthodes d'estimation utiles dans le cadre de ces modèles.

Le troisième chapitre présente un modèle de classification qui a pour objectif de séparer les effets liés à des facteurs exogènes communs des effets endogènes modélisés comme des processus stochastiques tout en identifiant des clusters de données. Cette méthode constitue l'un des principaux apports de ces travaux. Ce chapitre permet de présenter la formalisation du modèle et la méthode d'estimation utilisée pour estimer les paramètres du modèle. Ensuite, le modèle est évalué et comparé à d'autres modèles de référence basés sur des méthodes plus standard à partir d'un ensemble de jeux de données simulées. Pour finir, ce chapitre présente une extension du premier modèle, qui consiste à estimer les effets régressifs pour chaque cluster. La même démarche d'évaluation et de comparaison est alors menée.

Enfin, le quatrième chapitre est dédié à l'application des modèles sur des données simulées à partir d'un modèle thermique et de données réelles. En effet, dans un premier temps, un jeu de données de température intérieure pour un ensemble d'appartements est simulé à

partir d'un modèle de simulation thermique du bâtiment. Dans un second temps, les données issues d'une base de données open source REFIT sont utilisées afin de classifier des données de température issue d'un ensemble de maisons individuelles. Pour finir, des données de températures, de la base de données ANDRE, issues de parties communes de trois bâtiments, sont utilisées pour l'application du modèle de classification à centre dynamique et effets régressifs propres à chaque classe.

PUBLICATIONS

Publications dans des congrès

- Dynamic clustering and modeling of temporal data subject to common regressive effects, Louise Bonfils, Allou Samé, Latifa Oukhellou, *European Symposium On Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2021.
- Classification et décomposition de séries temporelles pour l'analyse de données thermiques de bâtiments, Louise Bonfils, Allou Samé, Latifa Oukhellou, *27ème rencontres de la Société francophone de Classification*, 2022.

Publications dans des revues

- Dynamic clustering and modeling of temporal data subject to common regressive effects, Louise Bonfils, Allou Samé, Latifa Oukhellou, *Neurocomputing, Volume 500, Pages 217-230*, 2022.

Chapitre 2

Contexte Énergie

Contents

2.1 Introduction

Dans le contexte actuel, la volonté de faire des économies d'énergie et de fournir cette énergie de manière plus durable sont des sujets importants et communs à de nombreux domaines, notamment celui de la construction et la rénovation de bâtiments, qu'ils soient résidentiels, industriels ou tertiaires. La modélisation thermique et physique permet de prédire la consommation d'un bâtiment en fonction de ses caractéristiques physiques (enveloppe, isolation, matériaux, système de chauffage...). Ces prédictions sont souvent éloignées de la consommation effective. L'une des raisons de ce décalage réside, dans le cas de bâtiment résidentiel, dans la mauvaise prise en compte et mauvaise connaissance des comportements des habitants des logements [De Wilde, 2014; Paone and Jean-Philippe, 2018].

C'est pour cette raison que l'extraction et la modélisation statistique de comportements à partir de données de consommation d'énergie ou de données thermiques constituent des domaines scientifiques majeurs dans le cadre de la prédiction de la consommation d'énergie de bâtiments. L'analyse des interactions des habitants dans et avec leur logement permet en particulier de :

- Réduire le biais entre les prédictions de la consommation d'énergie d'un bâtiment et les consommations effectives.
- Mettre en place des politiques de tarification pour orienter la consommation et éventuellement répartir la consommation nationale au cours d'une journée (lissage des pics).
- Intégrer ces comportements dans des modèles de simulation thermique du bâtiment et permettre d'améliorer le contrôle de certains équipements comme la VMC (Ventilation Mécanique Contrôlée) ou le chauffage collectif.

L'extraction des comportements peut passer par la classification des données liées à la consommation d'énergie de multiples bâtiments. Les travaux présentés dans ce chapitre s'inscrivent dans le cadre de la classification de données énergétiques. La prochaine section est donc destinée à présenter les travaux de classification qui ont pu être menés dans ce cadre applicatif. La seconde et troisième section de ce chapitre sont destinées à présenter deux principaux jeux de données réelles collectées au sein de logements et qui seront utilisés pour l'application des méthodes de classification développées dans le chapitre 4.

2.2 L'analyse de données dans le domaine de l'énergie : État de l'art

Comme cela a été évoqué dans le paragraphe précédent, l'analyse de données dans le domaine de l'énergie est utile à la modélisation et la prédiction de la consommation à l'échelle

d'un bâtiment, ou d'une ville par exemple. Par énergie, on entend l'électricité, le gaz ou encore le charbon et le bois selon les cas. Dans ce vaste domaine, l'analyse et la modélisation des comportements d'habitants au sein de logements sont des sujets d'intérêt. Il s'agit d'un terme large qui regroupe plusieurs types de comportements qui peuvent faire l'objet de modélisation ou d'analyses spécifiques. On peut parler de comportement de consommation d'électricité, de comportement de chauffage ou bien encore d'occupation. Les exemples sont nombreux et la première sous-section est destinée à définir la notion de comportement et comment les auteurs, qui ont analysé des données énergétiques, la définissent.

Un volet important des travaux réalisés dans le domaine de l'analyse de données énergétique étant la classification automatique de ces données dans le but d'extraire des comportements d'habitants au sein de bâtiments, la seconde sous-section est destinée à faire un état des lieux des méthodes et des travaux réalisés dans ce domaine. Il s'agit d'une des principales motivations aux travaux qui seront présentés dans les chapitres 4 et 5.

2.2.1 Le comportement des habitants dans des bâtiments résidentiels

Les habitants, par définition, vivent et interagissent avec le logement et le bâtiment dans lesquels ils se trouvent [Yan et al., 2015]. Par conséquent, certaines variables réagissent et sont impactées par ces comportements. En effet, on peut citer les éléments suivants :

- Les occupants d'un logement utilisent des services et des appareils qui consomment de l'électricité lors de leur utilisation (télévision, équipement de cuisine, machine à laver...).
- La présence et l'activité des occupants dans une pièce impactent les variables décrivant l'environnement de l'appartement comme la température intérieure, la concentration de CO2 ou encore l'humidité.
- L'interaction des occupants avec les fenêtres et les portes impacte également les conditions intérieures des logements.
- Les périodes de chauffe et la température de confort d'un ménage influencent la consommation d'énergie liée au chauffage ou à la climatisation.

Pour commencer, l'interaction des habitants avec les ouvrants est un comportement que des auteurs, comme dans [Rijal et al., 2007], cherchent à estimer dans le but de prédire la température intérieure d'un logement par exemple. De plus, tous les comportements qui impliquent une consommation d'électricité sont étudiés et sont sources d'intérêt. Selon l'objectif et le modèle utilisé, les auteurs peuvent s'intéresser à chaque comportement séparément dans le cadre des modèles "Bottom-up", qui consistent à déterminer les consommations électriques individuelles reliées à différentes activités (regarder la télévision, cuisiner...) afin d'estimer, dans un second temps, une consommation globale par agrégation [Widén and Wäckelgård, 2010; Diao et al., 2017]. Dans le cas de la classification de données énergétiques,

les courbes de charge globale sont souvent utilisées afin d'identifier des comportements de consommation type [Pan et al., 2017]. Dans ce cas, les auteurs considèrent les séries temporelles de consommation sur des périodes plus ou moins longues (données journalières, hebdomadaires, mensuelles) pour construire des profils type de consommation via des méthodes de classification. Dans le cas de données en grande dimension, c'est-à-dire avec un très grand nombre de données observées sur une longue période, les comportements peuvent être caractérisés par des variables représentatives extraites des courbes de charge [Beckel et al., 2012]. Ensuite, la demande de chauffage et les températures de confort peuvent également être considérées comme des comportements d'intérêt, car ils ont un impact fort sur les consommations liées au chauffage ou à la ventilation [Lévy et al., 2014; De Meester et al., 2013]. De plus, les comportements de chauffage, qui sont liés aux comportements de présence, font souvent l'objet de modélisation ou de classification [Lu et al., 2019].

Dans la littérature, l'occupation et l'activité dans les maisons sont fréquemment modélisées par des chaînes de Markov [Albert and Rajagopal, 2013; Delff Andersen et al., 2014]. En effet, ces modèles permettent d'inférer les périodes de présence et d'absence à partir de probabilités qui peuvent dépendre du temps. Cela permet de modéliser la présence sur une période choisie. L'occupation est généralement une composante des modèles "Bottom-up" qui constituent une famille importante dans le domaine de l'analyse de données énergétiques [Widén and Wäckelgård, 2010]. De plus, l'occupation ou les activités des habitants sont également des éléments importants dans l'estimation de la consommation d'énergie [Liisberg et al., 2016] ou des conditions intérieures [Wolf et al., 2019] via l'utilisation de modèles de Markov cachés. De plus, d'après [Amayri et al., 2019], l'estimation de l'occupation des logements pourrait permettre de mieux adapter la quantité d'eau stockée dans les ballons d'eau chaude pour les logements qui disposent de cet équipement. Ces études illustrent l'importance de l'occupation dans la prédiction de la consommation d'énergie et plus globalement dans l'analyse des comportements des occupants. Ces différents travaux motivent les modèles développés dans le chapitres 4 qui cherchent à identifier des comportements d'occupation des bâtiments.

Comme évoqué précédemment, l'analyse et l'estimation du comportement des habitants au sein d'un logement font l'objet de plusieurs travaux de recherche dont l'objectif est la prédiction de la consommation d'énergie. Les comportements, d'un appartement à l'autre ou d'une année à l'autre, peuvent être très variés et difficilement modélisables, car ils dépendent d'un nombre élevé de facteurs. Par conséquent, la classification, qui permet de regrouper des données en groupes homogènes, de les analyser, constitue une méthode appropriée dans ce cadre.

2.2.2 La classification de données énergétiques pour l'extraction des comportements d'habitants au sein de logements

La classification non supervisée est un sujet majeur dans la modélisation statistique et l'analyse de données. Elle est utilisée pour résumer l'information d'un ensemble d'observations, en un petit nombre de groupes. Les méthodes les plus utilisées sont les méthodes telles que les K-means ou les méthodes hiérarchiques (CAH) ([Wei et al., 2018]). Le chapitre 3 présente ces méthodes plus en détails. Dans le domaine de l'analyse de données d'énergie, la classification automatique qui a pour but d'identifier des groupes homogènes, permet de résumer l'information et d'extraire des comportements type relatifs aux habitudes de consommation, de chauffage, d'activité ou de présence par exemple. Il peut s'agir de comportements qui ne sont pas directement observables à partir des données récoltées. Par exemple, les mesures de CO2 au sein d'une pièce peuvent permettre d'inférer la présence des habitants [Franco and Leccese, 2020]. Dans le cadre de la classification, l'objectif est d'extraire des comportements homogènes à partir de données. Cette sous-section est dédiée à présenter un état de l'art de la classification de données de consommation d'énergie ou encore de données thermiques dans le but d'extraire des comportements type des habitants liés à leur consommation, à leur présence ou encore à leurs habitudes de chauffage.

La classification non supervisée de données d'énergie est un vaste sujet de recherche et il existe un grand nombre de méthodes pour procéder à la construction des clusters. L'une des méthodes qui est souvent utilisée est la méthode des K-means [Pan et al., 2017; Bourdeau et al., 2021]. Cette méthode est appréciée par sa simplicité de mise en œuvre et les bonnes performances qu'elle offre. Dans [Pan et al., 2017], les auteurs proposent une classification de courbes de charge journalières, issues d'un ensemble de logements représentatifs de la population étudiée via l'algorithme K-means. Une fois que les classes ont été déterminées ainsi que les courbes de charge type associées, les auteurs cherchent à décrire et à caractériser ces clusters. Pour cela, les clusters sont croisés avec des facteurs calendaires comme le type de jours et le mois de l'année, afin d'identifier à quels jours ou à quelle saison chaque cluster correspond. Cette seconde étape, qui vient après la classification, permet d'interpréter les clusters obtenus et les comportements de consommation qui y sont associés.

Dans ce dernier article, les auteurs utilisent les séries temporelles, représentant la consommation d'électricité horaire, comme données d'entrée. Cela peut poser un problème dans le cas de longues périodes observées ou de nombre important de séries disponibles. Pour réduire la quantité de données utilisées pour la classification, dans [Beckel et al., 2012], les auteurs utilisent un ensemble d'indicateurs calculés à partir des courbes de charges. Ces derniers proposent alors une classification de comportement de consommation d'électricité à partir de variables comme la consommation moyenne, la consommation maximale, la consommation minimale, la consommation moyenne au cours de la nuit ou d'un jour de week-end. Une fois la classification réalisée à partir de cet ensemble de variables caractérisant

les courbes de charges, les auteurs cherchent à caractériser les clusters en croisant les résultats avec des variables socio-démographiques disponibles pour décrire les ménages. Il en résulte que le profil de consommation est principalement lié à la catégorie socioprofessionnelle de la personne chef de famille, au nombre de chambres du logement ou encore la surface du logement. Cette seconde étape permet de caractériser les clusters et d'établir un lien entre les comportements de consommation type identifié par la classification et les caractéristiques socio-démographiques des ménages. Dans ce dernier article, les données sont classifiées à partir de la méthode Self-Organising Maps (SOM) basée sur des réseaux de neurones [Vesanto and Alhoniemi, 2000]. Il s'agit d'une méthode plus complexe, mais également souvent utilisée pour la classification de données énergétiques dans la mesure où elle est adaptée à des données en grande dimension. Dans [McLoughlin et al., 2015], les auteurs comparent les méthodes et concluent que la méthode Self-Organising Maps (SOM) est, dans le cas des données utilisées dans l'article, plus adaptée et offre de meilleures performances. Dans ce dernier article, les profils journaliers issus d'un ensemble de logements sont classifiés selon plusieurs méthodes. Comme pour les travaux de [Pan et al., 2017], les clusters regroupent des données journalières appartenant à plusieurs ménages. Néanmoins, les auteurs cherchent à caractériser les clusters selon les ménages qui les composent. Pour cela, les auteurs attribuent à chaque ménage le cluster le plus récurrent parmi l'ensemble des courbes journalières classifiées. Par la suite, des régressions logistiques sont estimées pour quantifier le lien entre le cluster d'appartenance et des caractéristiques socio-démographiques des ménages ou des caractéristiques des logements. Le choix d'utiliser des variables extraites des courbes de charge plutôt que les séries elles-mêmes se justifient par la taille des jeux de données. Ces questions ont été traitées dans [Yilmaz et al., 2019] où les auteurs comparent les résultats obtenus pour les deux situations. Dans le cas des données présentées dans l'article, l'utilisation des variables extraites permet d'obtenir des clusters plus compacts et distincts que l'utilisation des séries temporelles entières.

Les modèles de mélange constituent une famille importante de modèles pour la classification non supervisée. Les modèles qui seront présentés dans la suite des travaux appartiennent à la famille des modèles de mélange. Il est donc important ici d'aborder l'utilisation de ces modèles dans le contexte de données énergétiques. Dans [Stephen et al., 2014], les auteurs utilisent des modèles de mélange gaussien afin d'identifier des profils type de consommation d'électricité à partir de courbes de charge. Selon la période étudiée, la séquence de données observée peut-être longue et la classification doit alors être réalisée sur un grand nombre de données. Afin de réduire cette quantité de données, les auteurs utilisent des méthodes d'analyse factorielle pour classifier des données de dimension réduite. En effet, l'analyse factorielle permet de réduire le nombre d'observations utilisées pour la classification et ainsi éviter d'utiliser la séquence dans toute sa longueur. Une fois les classes de données identifiées, les comportements type de consommation associés à chaque cluster permettent de simuler ou de prédire la consommation énergétique d'un ensemble de ménages, d'estimer les pics

de consommation et le voltage minimum sur un réseau. Il est également possible d'utiliser directement les courbes de charge, sous forme de séries temporelles, pour construire des profils type de consommation, sans passer par une phase de réduction de dimensions. Dans [Melzi et al., 2017], les auteurs proposent une classification de courbes de charge journalière à partir d'un modèle de mélange gaussien auquel ils ajoutent des contraintes relatives au type de jour (Samedi, Dimanche, Semaine). Le modèle propose de classifier des courbes de charges en prenant en compte le type de jours comme une variable dépendante et ainsi construire des groupes d'occupants dont la consommation est similaire pour ces catégories de jours. Les clusters sont alors caractérisés par trois comportements type de consommation relatifs à ces trois types de jours. L'estimation des paramètres est réalisée via l'algorithme Expectation-Maximisation (EM). Une fois la classification réalisée, les auteurs cherchent à caractériser les clusters obtenus grâce à des données socio-démographiques et calendaires. Cette étape permet de déterminer les caractéristiques décrivant les occupants qui composent chaque cluster. Il est possible d'utiliser des modèles de mélange qui ne sont pas basés sur des distributions gaussiennes. Par exemple, dans [Granell et al., 2015] les auteurs proposent une méthode de classification de courbes de charge à partir de modèles de mélange de processus de Dirichlet. Ce choix de modélisation est justifié par l'utilisation d'une méthode appelée "Processus du restaurant chinois", qui permet d'estimer le nombre de clusters directement comme un paramètre au cours de la classification [Li et al., 2019]. Dans ce cas, les processus de Dirichlet sont utilisés comme les lois *a priori* des proportions de mélanges et le nombre de composantes du modèle n'est pas un hyperparamètre.

L'ensemble des travaux de classification présentés précédemment se basent sur les données de consommation énergétique. Il s'agit d'une partie importante des travaux de classification dans le domaine de l'énergie.

L'occupation des habitants au sein des logements et des bureaux est également un sujet d'intérêt important. En effet, la prédiction de l'occupation au sein d'un logement peut permettre, par exemple, de mieux contrôler le chauffage en faisant correspondre les périodes de chauffe avec les périodes de présence. Des auteurs, comme [Vázquez and Kastner, 2011], cherchent à extraire des comportements d'occupation type à partir de données de présence au sein d'un ensemble de logements. L'objectif de ces travaux est de prédire l'occupation afin de contrôler le chauffage et d'améliorer le confort des habitants. La classification est réalisée à partir de plusieurs méthodes (K-means, SOM, Fuzzy C-means...) qui sont comparées à l'aide d'un ensemble d'indicateurs. Dans [D'Oca and Hong, 2015], les auteurs cherchent à identifier des patterns type d'occupation dans des bureaux à partir d'une méthode de classification K-means sur des données de présence.

Les travaux de classification de données énergétiques présentés précédemment ont pour objet d'étude, en grande partie, des séries temporelles et plus particulièrement des courbes de charge. Cependant, la modélisation des centres de classes ne permet pas directement de prédire les valeurs futures ou encore d'estimer la dynamique de ces centres. Pourtant,

on peut imaginer que la mise en place de politiques incitatives, de changements de prix ou d'innovations peuvent entraîner des modifications des comportements et habitudes. Il peut donc être intéressant de considérer l'aspect dynamique et évolutif du comportement dans la tâche de classification, en modélisant les paramètres des classes comme des processus stochastiques par exemple. L'évolution des comportements dans les problèmes de clustering est souvent prise en compte en utilisant des méthodes de segmentation pour identifier les périodes au cours desquelles les comportements sont statiques et constants, puis effectuer la classification, sur ces périodes spécifiques. Par exemple, dans [Liisberg et al., 2016], les auteurs cherchent à identifier des profils type d'occupants à partir d'une méthode de classification indirecte basée sur la modélisation de la présence des habitants à partir de chaînes de Markov. En effet, à partir de courbes de charge journalière, la méthode consiste d'abord à identifier des états liés à la présence, à l'aide d'un modèle de Markov caché inhomogène. Cette première étape permet d'associer une séquence de probabilité pour chaque état identifié à chaque courbe journalière. Ces séquences de probabilités sont ensuite classifiées afin d'extraire des comportements de présence type au sein d'appartements. Ce choix de modélisation pourrait permettre de prédire la présence au sein d'un logement par exemple. La modélisation de la présence à partir de chaînes de Markov se retrouve également dans les travaux de [Widén and Wäckelgård, 2010] et [Delff Andersen et al., 2014], où les auteurs modélisent les séquences de présence, respectivement au sein de logements et de bureaux, à partir de chaînes de Markov inhomogènes. Cela permet d'estimer des probabilités de transition de l'état présent à l'état absent. Ces probabilités sont intéressantes dans la mesure où elles permettent d'estimer la présence à l'échelle d'un bâtiment entier ou encore de servir d'input dans des modèles de prédiction de consommation de type "Bottom-up" [Widén and Wäckelgård, 2010; Diao et al., 2017].

Dans le but de développer un modèle de classification de séries temporelles, [Devijver et al., 2015] proposent une décomposition en ondelettes des courbes de charge pour réduire la dimension des données d'entrée et sélectionner les éléments des séries qui sont considérés comme les plus discriminants pour la classification. Les travaux cités précédemment ont tous en commun de chercher à modéliser, à l'aide de processus temporels, des comportements extraits à partir de données énergétiques telles que des courbes de charges ou encore des variables caractéristique de la consommation des occupants. Les travaux présentés par la suite s'inscrivent dans ce cadre, avec le développement d'une méthode de classification qui cherche à caractériser les clusters à partir de processus stochastiques.

2.2.3 L'impact de données contextuelles sur les comportements dans le cadre de la classification de données énergétiques

La température ambiante d'une maison est influencée par l'état du bâtiment, mais aussi par les conditions météorologiques extérieures ainsi que par l'occupation et l'activité des

habitants.

Ainsi, l'extraction et l'analyse des comportements des habitants doivent tenir compte de facteurs contextuels comme le type de jours, la météo extérieure ou le mois de l'année. En d'autres termes, l'extraction des comportements peut être couplée avec l'identification d'effets liés à des variables contextuelles afin de mettre en évidence l'impact de variable météorologique ou calendaires sur les comportements identifiés. Les auteurs de [Kane et al., 2017] utilisent la température extérieure et la température intérieure pour calculer et estimer le comportement de chauffage dans un ménage anglais typique. On peut également citer [Pardo et al., 2002] ou encore [Fazeli et al., 2016], dans lesquels les auteurs cherchent à modéliser le lien entre la température extérieure et la demande de consommation électrique. Dans les travaux de [Widén and Wäckelgård, 2010], les auteurs proposent de modéliser la présence à partir d'une chaîne de Markov inhomogène en incluant la température extérieure comme covariable. Ces travaux montrent l'importance de la prise en compte du contexte extérieur pour l'extraction de comportements au sein des logements. Le type de jours et le jour de la semaine sont également des facteurs exogènes considérés comme ayant un impact sur les comportements. En effet, dans [Pan et al., 2017], les auteurs pointent le fait que les comportements sont différents entre la semaine et le week-end. Le mois de l'année ou la saison sont également considérés comme des facteurs influençant le comportement des occupants [Pan et al., 2017; Diao et al., 2017]. Les travaux cités précédemment établissent le lien entre les facteurs de contexte et les comportements une fois la classification a été réalisée. On peut imaginer utiliser ces facteurs lors de la classification. Dans [Melzi et al., 2017], les auteurs construisent des classes de comportements de consommation pour différents types de jours. Cela permet de discriminer d'abord les données afin que les comportements type identifiés, pour chaque type de jours, ne soient pas liés à ces facteurs contextuels. Dans le cadre des travaux présentés par la suite, ces travaux sont importants, car ils justifient d'une part l'effet de variables météorologiques et calendaires sur les comportements des occupants et d'autre part l'intérêt d'utiliser ces données contextuelles au cours de l'étape de classification.

2.2.4 Positionnement du travail par rapport à l'état de l'art

Les travaux présentés dans les chapitres 4 et 5 s'inscrivent dans le cadre de la classification de données énergétiques et thermiques dans le but d'extraire des comportements d'occupants inobservés relatifs aux habitudes de consommation, d'occupation ou de chauffage.

Les comportements d'occupants sont très variés, car ils dépendent d'une multitude de facteurs qu'il est souvent difficile d'identifier ou d'observer. Par conséquent, selon la zone géographique, le type de logement, le type de bâtiment ou encore la composition du ménage, les comportements de consommation ou les habitudes de chauffage peuvent être très différents. Il n'est toutefois pas toujours facile de disposer de données relatives à ces

comportements. Il est donc important de construire des modèles qui permettent d'identifier des profils type à partir de données facilement mesurables et observables. On a également souligné l'importance de l'occupation et de la présence sur le niveau de consommation ou sur le chauffage d'un logement. Par conséquent, on a la volonté d'analyser et d'identifier des comportements relatifs à l'occupation et à l'activité des occupants.

Dans le cadre de la classification de séries temporelles, certains auteurs cherchent à modéliser les comportements à partir de processus dynamiques ([Devijver et al., 2015; Granell et al., 2015]. Cela permet de modéliser l'évolution des comportements au cours du temps et éventuellement de prédire les futurs comportements. Les méthodes développées dans le chapitre 4 s'inscriront dans ce cadre, avec des profils de classes considérés comme des processus autorégressifs.

Un autre atout des travaux présentés par la suite réside dans la prise en compte d'effets régressifs liés à des facteurs météorologiques ou calendaires lors de la classification. Cette volonté est motivée par de nombreux travaux concernant la classification, qui montrent le lien entre les comportements identifiés et des variables calendaires comme le type de jours ou les conditions météorologiques. Par conséquent, on souhaite identifier, dès la phase de classification, l'effet de ces facteurs exogènes et contextuels afin d'identifier des comportements uniquement liés à des variables inobservées comme la présence ou les températures de confort.

Les modèles construits dans le chapitre 4 se placent dans le cadre de la classification de données dans le but d'extraire des comportements d'habitants relatifs à leur présence ou encore leur comportement de chauffage. Le chapitre 5 est dédié à l'application de ces modèles sur des jeux de données thermiques qui sont présenté dans la section suivante.

2.3 Le jeu de données réelles REFIT

Le jeu de données REFIT, disponible en accès libre [Firth et al., 2017], est constitué de données énergétiques et thermiques récoltées au sein de 20 maisons individuelles en Angleterre. Le projet REFIT s'est déroulé de mai 2012 à octobre 2015. Il est réalisé, en collaboration, par l'université de Loughborough, l'université de Strathclyde et l'université d'East Anglia. Il a été lancé afin d'étudier, d'analyser et de promouvoir à long terme les maisons dites "intelligentes", équipées d'un ensemble de capteurs et ayant accès à des services pour mieux contrôler, superviser et réguler la consommation énergétique, le chauffage, la qualité de l'air ou encore la fermeture et l'ouverture des volets par exemple. L'objectif du projet REFIT est de promouvoir les maisons équipées de dispositifs de mesure et de suivi des consommations, du chauffage et de l'activité, en étudiant de quelle manière le développement de tels équipements peut permettre de réduire la demande en énergie des ménages. Dans le cadre de ce projet, 20 maisons ont été équipées de capteurs pour mesurer :

- la température dans différentes pièces de la maison,
- l'humidité dans différentes pièces de la maison,
- la consommation d'électricité de différents équipements,
- la consommation d'électricité globale de la maison,
- la consommation de gaz de la maison,
- les mouvements détectés dans la pièce de vie de la maison,
- les conditions météorologiques du quartier regroupant les maisons,
- des variables catégorielles descriptives de la maison et des habitants.

Au-delà des variables citées ci-dessus, la base de données REFIT contient des données issues d'entretiens semi-dirigés réalisés auprès des ménages habitant les maisons par rapport à leur ressenti vis-à-vis des maisons intelligentes. De plus, une enquête nationale sur la perception des maisons intelligentes a été réalisée auprès de plus de mille propriétaires. Enfin, la base de données regroupe des données de température, de surface de radiateur ou de murs pour certaines maisons.

Cette section est dédiée à la présentation des données REFIT et plus particulièrement des variables qui seront utilisées et abordées par la suite. Ces variables d'intérêt regroupent les données de température, la consommation d'électricité, les conditions météorologiques et les données issues des détecteurs de mouvements. Les variables catégorielles auraient pu être une source intéressante d'information. Cependant, le nombre restreint de maisons et le nombre relativement important de données manquantes rendent l'utilisation de ces variables dans l'analyse des classes difficiles.

2.3.1 Les caractéristiques des maisons et des habitants

Le jeu de données REFIT est composé de variables caractéristiques sur les ménages et leur maison. Le tableau 2.1 décrit les données disponibles. Les maisons sont caractérisées notamment par l'épaisseur de l'isolation, l'année de construction ou encore la surface et le volume de la pièce de vie. Les ménages, qui habitent les maisons, sont caractérisés par leur taille et leur composition ainsi que l'activité du chef de famille.

On se focalise sur les variables continues disponibles dans le jeu de données REFIT à savoir des données de consommation d'énergie des maisons, des données de température, de présences et des données météorologiques. L'objectif de cette sous-section est de présenter les données collectées et disponibles et éventuellement les premiers retraitements réalisés. On commencera d'abord par les données de température, puis les données de la consommation d'électricité, ensuite les données issues des détecteurs de mouvements et pour finir les variables météorologiques.

TABLE 2.1 – Variables, de la base REFIT, caractérisant les maisons et leurs habitants.

	Composition des ménages		Caractéristiques des maisons				
	Nombre d'habitants (Adultes + Enfants)	Activité du chef de famille	Année de construction	Type d'isolation	Epaisseur de l'isolant (mm)	Surface de la pièce de vie (m^2)	Volume de la pièce de vie (m^3)
M01	2 (2 + 0)	Employé à plein temps	1975-1980	Laine/ Fibre de verre	300	23,76	56,60
M02	4 (2 + 2)	Employé à plein temps	1919-1944	Laine/ Fibre de verre	300	18,00	43,30
M03	2 (2 + 0)	Employé à mi-temps	1981-1990	Laine/ Fibre de verre	300	25,76	61,82
M04	2 (2 + 0)	Retraité	1850-1899	Laine/ Fibre de verre	>300	31,13	85,60
M05	4 (2 + 2)	Employé à plein temps	1850-1899	Laine/ Fibre de verre	300	17,20	44,72
M06	2 (2 + 0)	Employé à mi-temps	ap 2002	Laine/ Fibre de verre	300	19,35	46,44
M07	4 (2 + 2)	Employé à plein temps	1965-1974	Mousse rigide	50	20,19	48,45
M08	2 (2 + 0)	Retraité	1965-1974	Laine/ Fibre de verre	>300	24,94	59,86
M09	2 (2 + 0)	Employé à plein temps	1919-1944	Laine/ Fibre de verre	50	17,39	43,48
M10	4 (2 + 2)	Employé à plein temps	1919-1944	Billes de vermiculite	100	18,90	47,25
M11	1 (1 + 0)	Retraité	1945-1964	Laine/ Fibre de verre	200	20,52	47,20
M12	2 (2 + 0)	Employé à plein temps	1991-1995	Laine/ Fibre de verre	>300	16,12	37,08
M13	4 (2 + 2)	Employé à plein temps	ap 2002	Laine/ Fibre de verre	>300	24,19	60,46
M14	1 (1 + 0)	Employé à plein temps	1965-1974	Laine/ Fibre de verre	>300	16,63	39,91
M15	6 (2 + 4)	Employé à plein temps	1981-1990	Inconnue	25	25,55	58,76
M16	3 (2 + 1)	Employé à plein temps	1965-1974	Laine/ Fibre de verre	>300	22,10	50,83
M17	2 (2 + 0)	Retraité	1965-1974	Laine/ Fibre de verre	>300	30,74	73,78
M18	4 (2 + 2)	Employé à plein temps	1945-1964	Couette haute performance	200	20,25	48,60
M19	2 (2 + 0)	Employé à plein temps	1965-1974	Laine/ Fibre de verre	250	16,77	37,73
M20	4 (2 + 2)	Employé à plein temps	1981-1990	Laine/ Fibre de verre	300	17,92	43,01

2.3.2 Données de température intérieure

La première variable d'intérêt est la température intérieure. Le modèle proposé dans le chapitre 4 se base sur ces données mesurées au sein des logements pour identifier des comportements d'occupation notamment. Le chapitre 5, dédié à l'application de ces modèles à des données thermiques, se base en partie sur les données de température du jeu de données REFIT.

La collecte des données, réalisée au sein de 20 maisons, commence, au plus tôt, le 18 septembre 2013 et se termine, au plus tard, le 14 août 2015. Les données sont collectées dans les pièces à vivre des maisons (living-room), au pas de temps 30 minutes pour la plupart. Pour certaines maisons (maison n°1, n°7 et n°13), les données de température sont disponibles au pas de temps 15 minutes à partir de février 2015. La raison de ce changement n'est pas détaillée, mais provient probablement du remplacement des capteurs. Afin d'harmoniser

les données, pour les maisons et la période concernée, les mesures ont été agrégées au pas de temps 30 minutes en conservant la température moyenne. Le graphique 2.1 représente les données manquantes pour les capteurs de température des pièces de vie des 20 maisons. On peut voir que les périodes de collecte ne sont pas tout à fait les mêmes, et que pour certaines maisons (maison n°2 et maison n°17), la température n'a pas été mesurée pendant une période assez longue. Par conséquent, l'application des modèles de classification devra être réalisée sur des périodes réduites pour utiliser les données de l'ensemble des maisons.

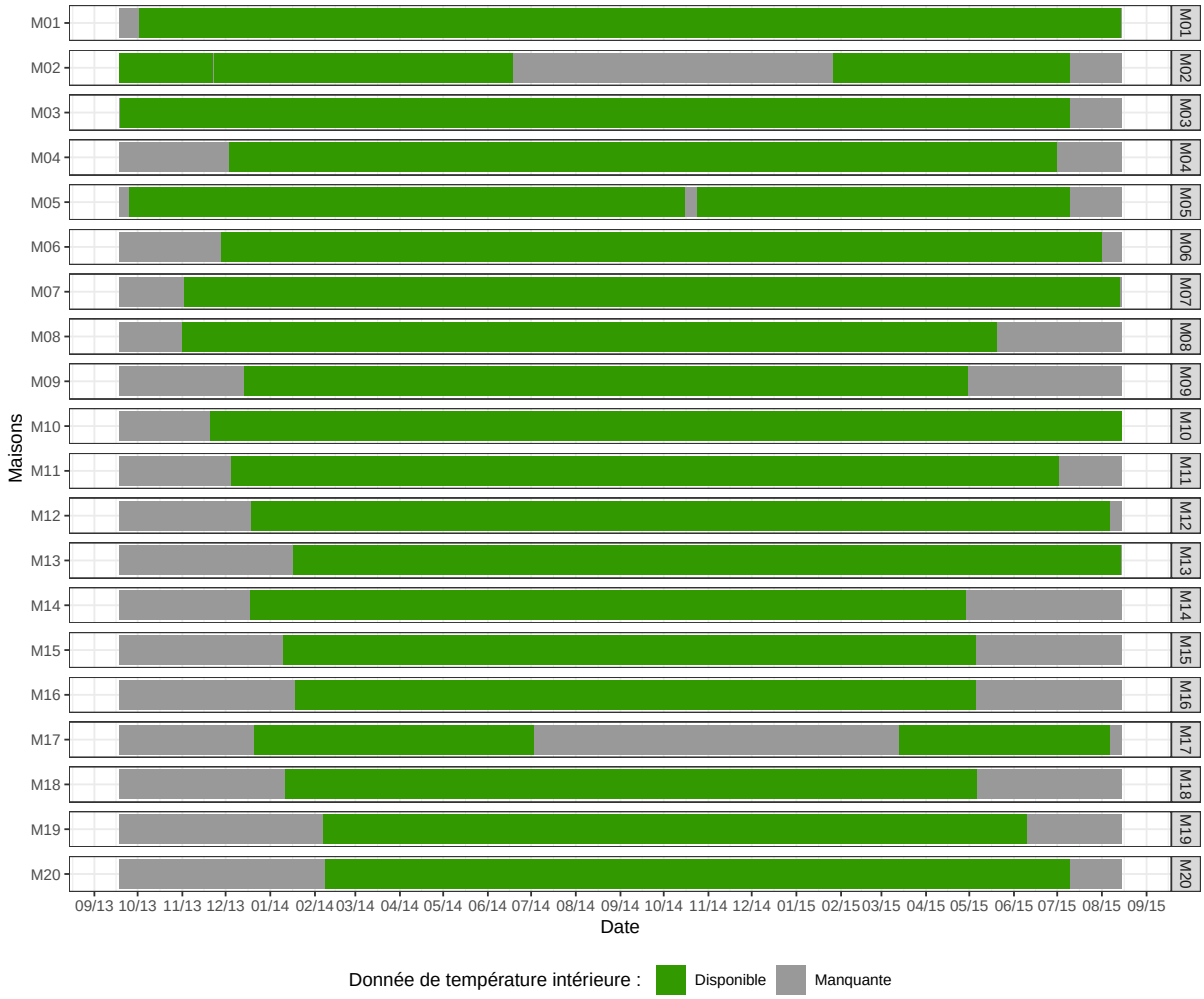

FIGURE 2.1 – Graphique représentant les données manquantes pour les capteurs de température dans les pièces de vie des 20 maisons. Les données sont collectées ou agrégées au pas de temps 30 minutes.

Les données sont manquantes pour de longues périodes différentes selon les maisons. Afin d'utiliser le plus de données, et donc de maisons possibles, il faudra, par la suite, se baser sur des périodes au cours desquelles les mesures ne sont pas manquantes pour un maximum de maison. Les périodes utilisées seront détaillées lors de l'application dans le chapitre 5.

Le graphique 2.2 représente les températures intérieures moyennes mesurées au sein de chaque maison au cours des différents mois de l'année. On remarque qu'une certaine tendance se dessine pour chaque mois, ce qui montre que les données de température

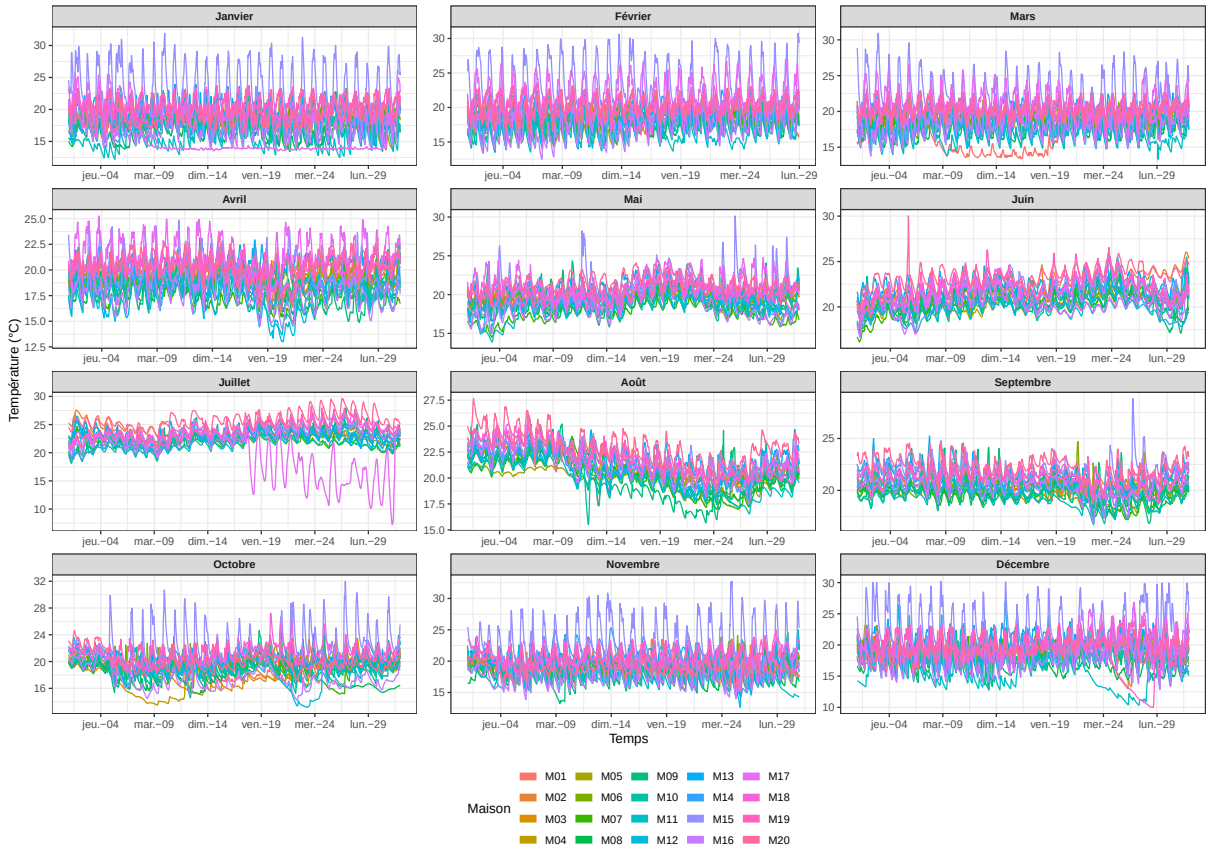

FIGURE 2.2 – Températures intérieures moyennes, au pas de temps 30 minutes, au cours de chaque mois de l'année pour les différentes maisons de la base REFIT.

intérieure sont bien impactées par des facteurs calendaires par exemple, mais on remarque aussi des variations propres à certaines maisons. Ces variations sont liées au logement et aux comportements des habitants de celui-ci. On observe également moins de variabilité entre les maisons au cours de l'été comparé à la période de chauffe allant d'octobre à avril.

Ensuite, le graphique 2.3 représente les températures moyennes au cours d'une semaine pour chaque maison. À une échelle temporelle plus petite, on peut également observer un pattern général similaire périodique, mais avec des différences sur l'amplitude de la température et sur la durée des périodes avec une température basse et haute.

Ces deux graphiques permettent de motiver l'utilisation des variables décrivant l'environnement intérieur des logements afin d'analyser et de classifier des comportements relatifs à des habitudes de chauffage, d'occupation et d'activité dans un logement.

La température intérieure semble donc être influencée par des variables saisonnières ou encore le moment de la journée, bien qu'on observe tout de même des différences d'une maison à l'autre qui pourrait correspondre à des comportements différents au sein du logement. Dans la même lignée, l'humidité intérieure pourrait également donner des informations sur les comportements des habitants dans leur logement.

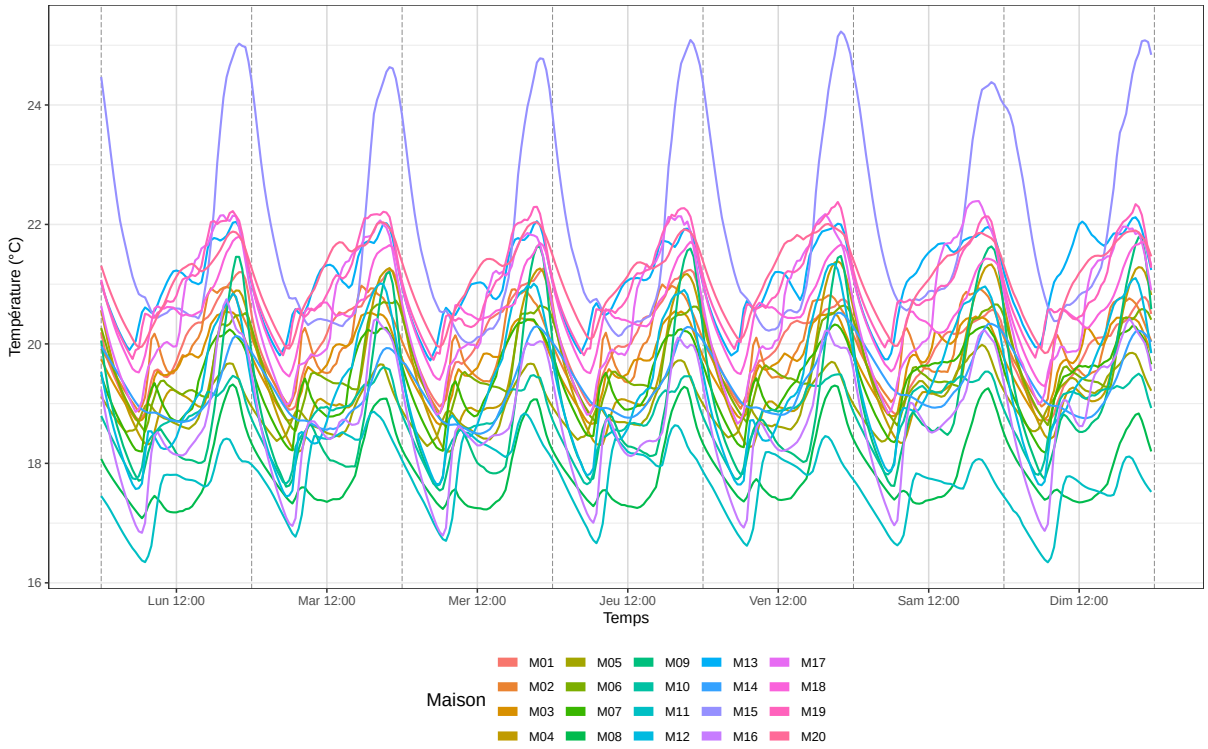

FIGURE 2.3 – Courbes hebdomadaires moyennes de la température intérieure pour les logements de la base REFIT, au pas de temps 30 minutes.

2.3.3 Données de consommation d'électricité

La consommation d'électricité est mesurée, pour l'ensemble des 20 maisons sauf la maison n°14, mais une 21^{e} maison est ajoutée au jeu de données. Les mesures ont commencé, au plus tôt, le 17 septembre 2013 et se sont terminées, au plus tard, le 10 juillet 2015. Les données d'électricité sont initialement récoltées au pas de temps 1 minute. Il est important de noter que les maisons disposent d'un système de chauffage au gaz. La consommation d'électricité est donc le résultat de l'utilisation des différents appareils de la maison, de la lumière uniquement ou éventuellement d'un chauffage d'appoint.

Comme on peut le voir sur le graphique 2.4, il arrive que les données ne soient pas toujours disponibles à chaque minute et que lorsqu'elles sont manquantes, elles le sont pendant une période d'au moins 30 minutes. De plus, les autres variables sont récoltées à un pas de temps plus important (15 ou 30 minutes). Cela simplifierait le traitement de cette variable si la consommation d'électricité était agrégée à la demi-heure. Pour cela, il y a deux possibilités. Il semblerait plus logique, au vu de la nature de la variable, d'agréger les données en conservant la quantité totale d'électricité consommée par tranche de demi-heure. Cependant, il faudrait, dans un premier temps, interpoler les données puis les agréger. La seconde option, qui est retenue, est d'agréger les données avec la consommation moyenne d'électricité par demi-heure. Ce choix ne devrait pas avoir d'impact sur l'analyse des variations observées et des dynamiques de consommation.

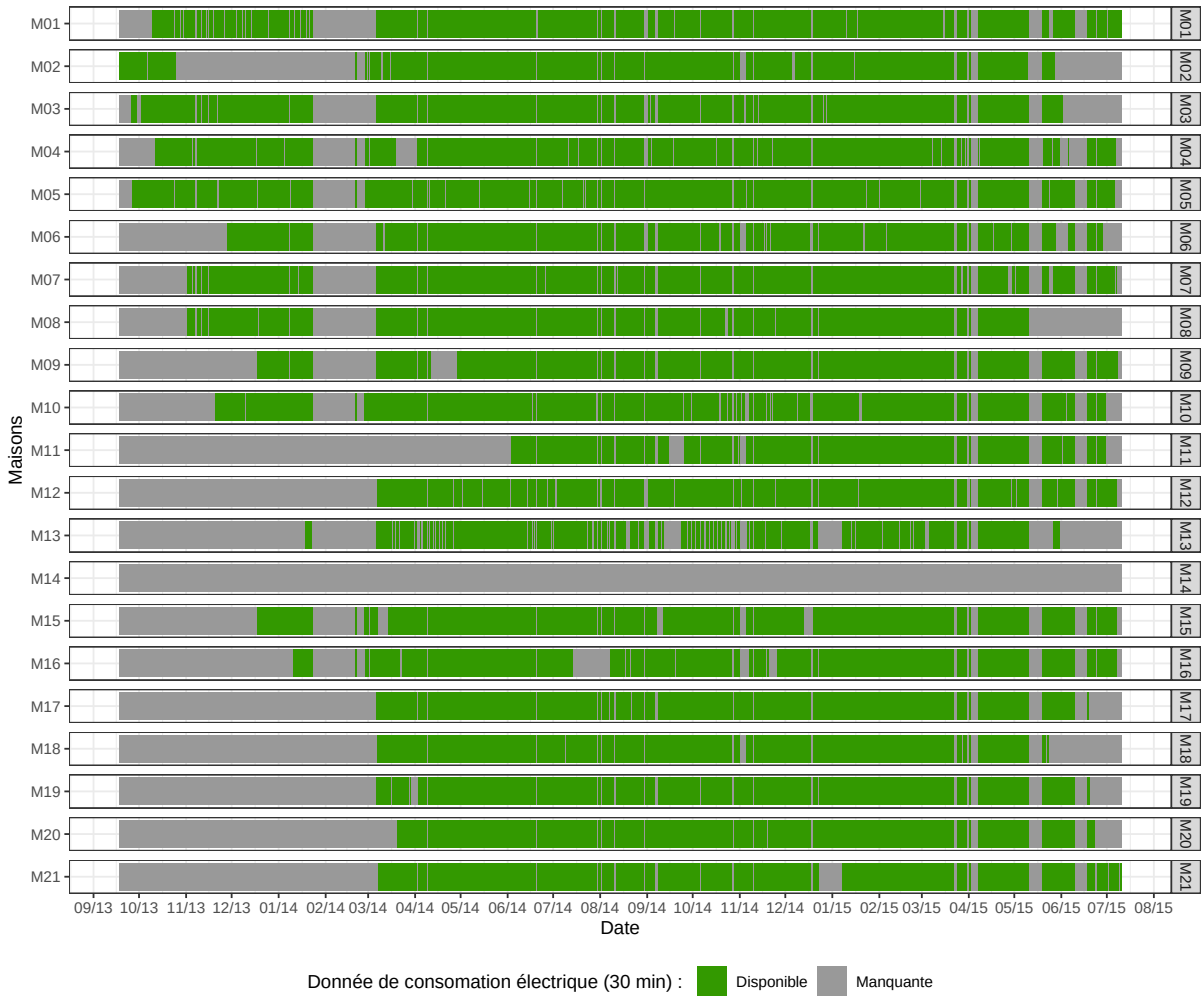

FIGURE 2.4 – Données manquantes pour les capteurs de la consommation électrique. Le graphique représente le nombre de mesures, normalement réalisée toutes les minutes, manquantes par tranches de 30 minutes.

INTERPOLATION

Les données vont être utilisées sur des périodes plus ou moins longues (entre 1 semaine et 1 mois). Par conséquent, si au cours de la semaine, certaines mesures sont manquantes, la perte d'information reste faible, mais l'estimation est rendue difficile. Il faut donc s'assurer d'utiliser, en entrée des modèles qui seront par la suite présentés, uniquement les données qui ne contiennent aucune longue période de valeur manquante. Par conséquent, pour les longues périodes (supérieure à un jour) de données manquantes, la période devra, dans la mesure du possible, ne pas être considérée. Par contre, pour les courtes périodes de données manquantes (quelques heures), il est possible d'interpoler les données. La méthode qui a été choisie ici est l'interpolation linéaire. C'est-à-dire, qu'on considère qu'au cours d'une période de données manquantes, la consommation électrique a évolué linéairement entre deux les instants où les mesures sont disponibles.

Le graphique 2.5 représente la consommation journalière moyenne selon le type de jour de la semaine. Ce graphique montre que, malgré une tendance commune (pic de consommation

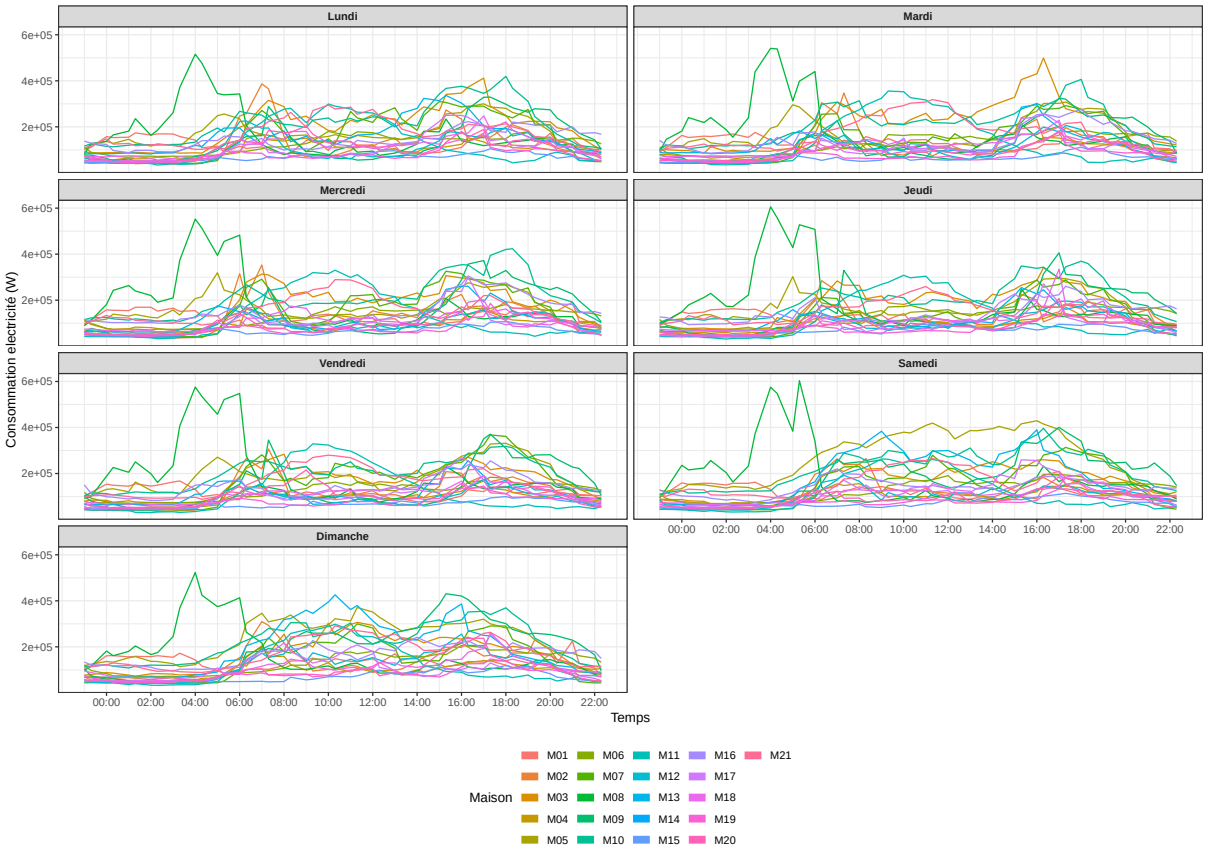

FIGURE 2.5 – Courbes journalières moyennes de consommation d'électricité par type de jour pour les logements de la base REFIT.

le matin et l'après-midi), certaines maisons se distinguent par des pics plus tôt (maison n°8), des pics plus ou moins longs (maison n°9) ou l'absence de pics certains jours (maison n°5 le samedi). De plus, on peut également voir que les comportements sont légèrement différents entre les jours de la semaine et du week-end. Ces observations peuvent justifier d'inclure des variables calendaires comme le jour de la semaine ou le type de jours dans les variables exogènes qu'on considère comme ayant des effets sur les données de consommation. Le second graphique 2.6 permet d'observer la périodicité dans la consommation d'énergie ainsi que des différences d'amplitude et de durée des pics pour les différentes maisons.

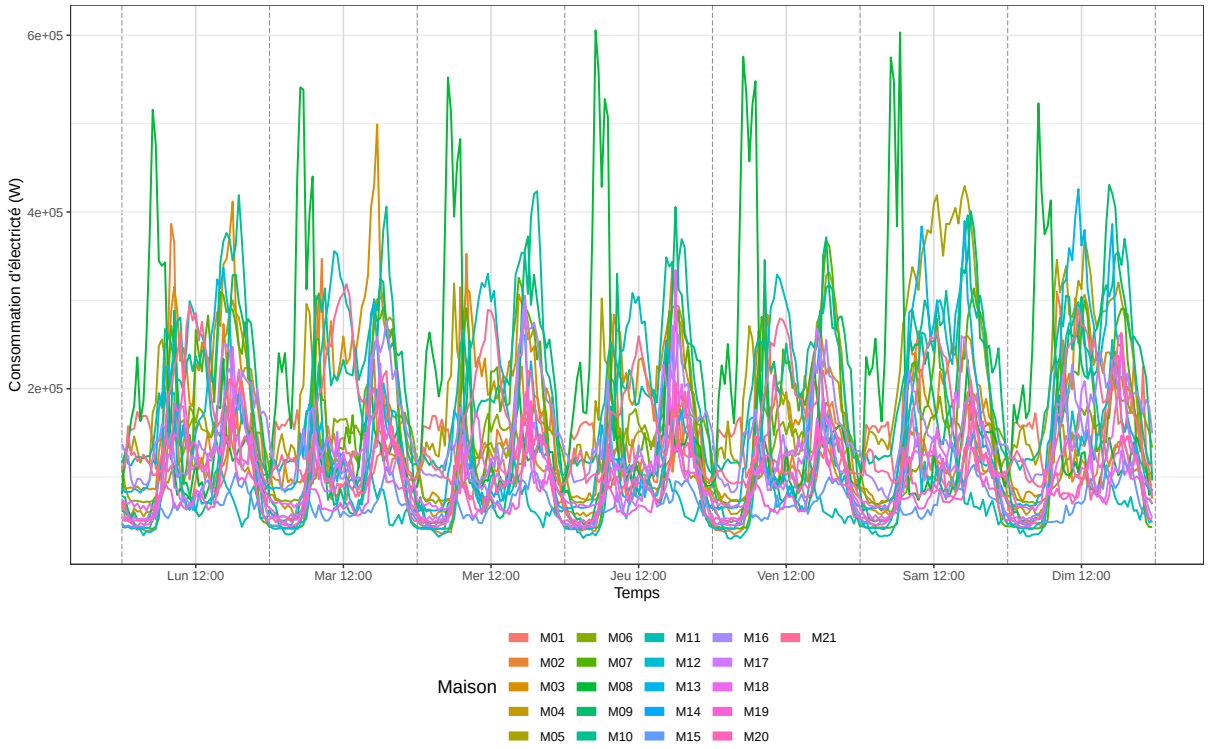

FIGURE 2.6 – Courbes hebdomadaires moyennes de consommation d'électricité pour les logements de la base REFIT.

2.3.4 Données de présence : détecteur de mouvement

Le projet REFIT, dans un but d'analyser des interactions entre les habitants et les différents services et équipements de leur maison, avait prévu des capteurs de mouvements au sein des maisons. Ainsi, on dispose de données de détecteurs de mouvements dans les pièces de vie de 19 maisons sur les 20 initiales (il manque la maison n°8). Ces mesures ont été réalisées à partir du 13 juin 2014 jusqu'au 28 avril 2015.

Les données de détecteurs de mouvements sont collectées toutes les minutes, sous la forme d'une variable binaire, indiquant si un mouvement a été détecté ou non. On souhaite transformer cette variable binaire en variable de comptage au pas de temps 30 minutes, pour coïncider avec les autres variables (température, consommation d'électricité...). Cette nouvelle variable de comptage est obtenue en sommant les données binaires par tranche de 30 minutes. Le nombre de détections de mouvement permet d'inférer l'occupation et l'activité des habitants au sein des pièces de vie. Comme pour les autres variables, les périodes de collecte ne sont pas les mêmes pour les différentes maisons. Le graphique 2.7 représente les données manquantes et disponibles.

Comme pour les précédentes variables, le choix de la période considérée pour utiliser ces données devra permettre de perdre un minimum de données et de maisons.

Le graphique 2.8 représente le nombre journalier moyen de détections selon le jour de la semaine. Comme pour les autres variables, on observe une périodicité au cours de la semaine

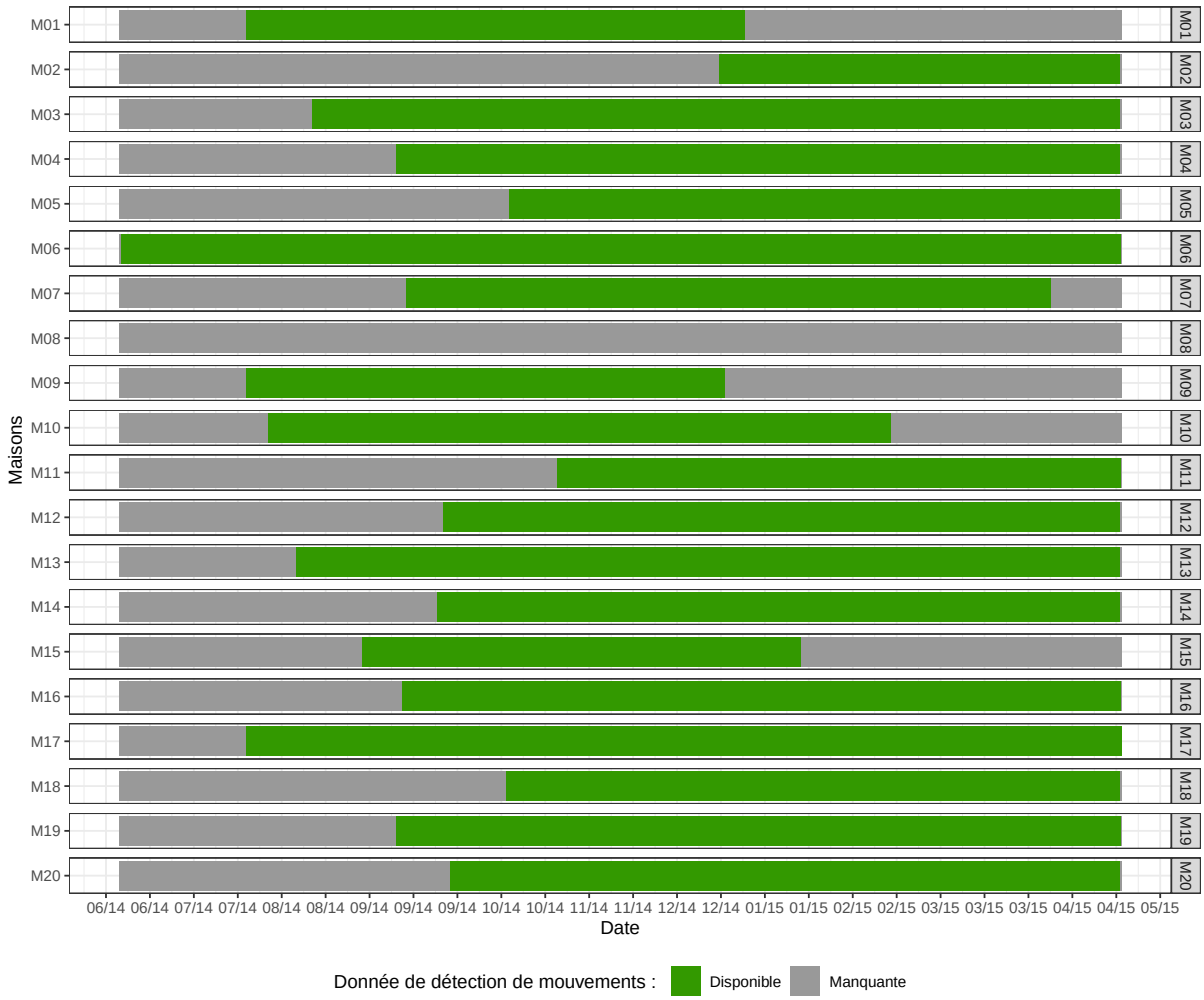

FIGURE 2.7 – Données manquantes pour les données de comptage de détection de mouvement au sein des pièces de vie des maisons.

avec des subtilités propres à chaque maison. De plus, sur une journée, on remarque des différences selon le type de jours et une tendance générale avec deux pics d'activité, mais là encore avec certaines exceptions.

Finalement, la base de données regroupe des données de température intérieure, de consommation d'électricité et de détections de mouvements. On peut observer que les courbes journalières moyennes de nombre de détections de mouvement et de la consommation d'électricité sont assez similaires. Ces deux variables sont liées aux comportements d'occupation des habitants au sein des logements. Les travaux présentés dans le chapitre 4 et 5 cherchent à extraire de tels comportements à partir des données de températures intérieures.

Comme on l'a dit précédemment, certaines données semblent être impactées par la saison, le jour de la semaine ou l'heure de la journée. On peut également imaginer que des variables météorologiques pourraient aussi impacter la température intérieure ou la présence notamment.

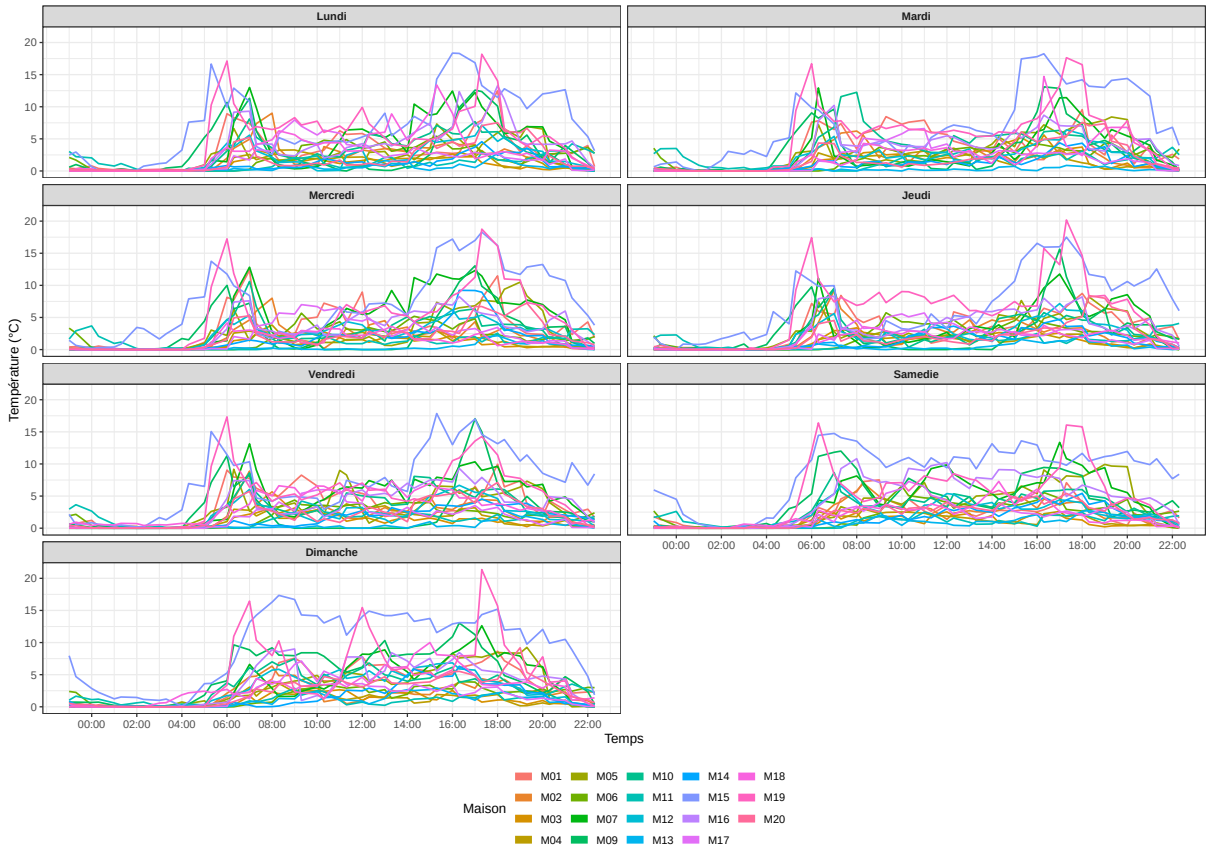

FIGURE 2.8 – Nombre de détections de mouvement moyen au cours d'une journée, pour chaque jour de la semaine et pour chaque maison.

2.3.5 Données météorologiques : température, humidité, volume de précipitations

Les données météorologiques sont mesurées et collectées par une station située à proximité du quartier regroupant les maisons. La station météo permet de mesurer la température extérieure (°C), l'humidité (%), l'irradiance solaire (MJ/m²), les précipitations (mm) et la vitesse du vent (m/s). Ces données seront utiles lors de l'application du chapitre 5, car elles constituent des variables de contexte aux données décrivant l'environnement des logements. Les données ont été collectées entre le 4 novembre 2013 et le 1er juin 2015, au pas de temps 15 minutes, sans aucune donnée manquante. Afin d'uniformiser le jeu de données, les mesures ont été agrégées au pas de temps 30 minutes. Ces données météorologiques permettent d'avoir un contexte commun à l'ensemble des mesures réalisées dans les maisons. On peut alors isoler les effets de ces variables sur les maisons pour ensuite se concentrer sur la partie restante correspondant aux comportements individuels des habitants. C'est le sujet et l'objectif des modèles qui seront présentés dans le chapitre 4, puis appliqués aux données que l'on vient de présenter, dans le chapitre 5.

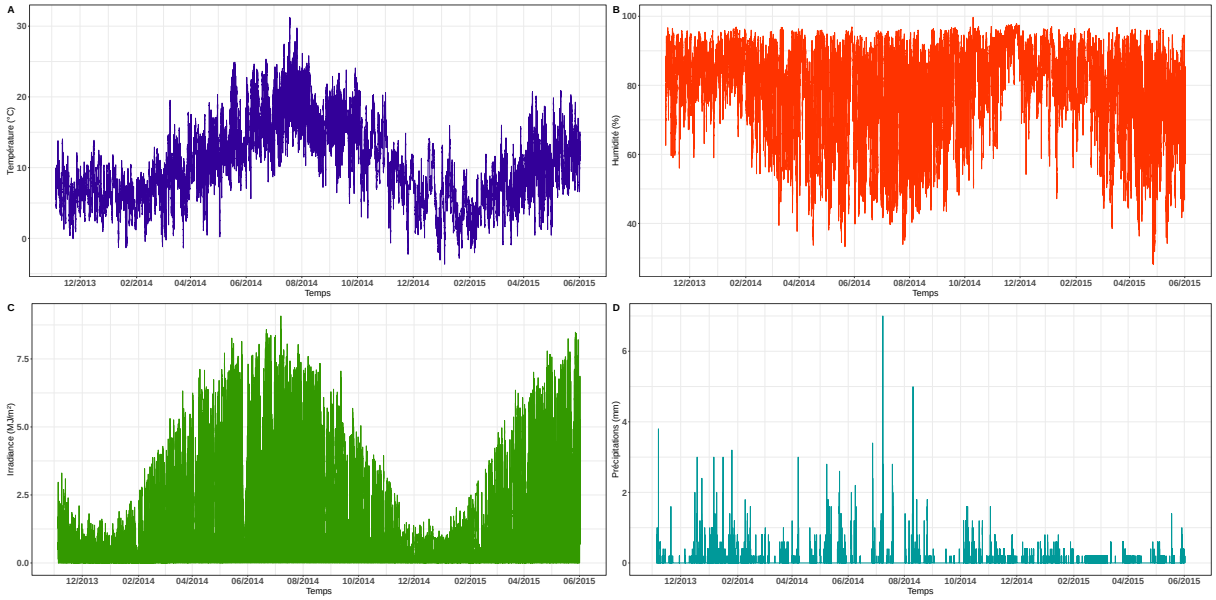

FIGURE 2.9 – Données météorologiques sur l'ensemble de la période d'observation pour la base REFIT. (A) représente la température extérieure, (B) représente l'humidité extérieure, (C) représente l'irradiance solaire et (D) représente les précipitations horaires moyennes.

2.4 Le jeu de données réelles ANDRE

Le second jeu de données disponibles pour les applications est issu du projet I-site Future ANDRE. Le projet "Modèle pour l'**AN**alyse, la Décomposition et le **R**econstruction de données de consommations **E**nergétiques" est un projet qui s'articule autour de trois grands volets et qui a pour objectif principal une meilleure connaissance et compréhension des comportements des habitants et de leur interaction avec les bâtiments afin d'améliorer la prédiction de la consommation d'énergie d'un bâtiment. Les travaux de cette thèse ont été initiés par ce projet dans le but de développer des modèles de classification applicables à des données récoltées au sein d'appartements d'un parc de logements sociaux dans l'Est de l'Ile de France. À l'heure actuelle, seules les données des parties communes des bâtiments sont disponibles et ont été traitées.

Le projet ANDRE s'organise en trois volets : un volet physique, qui a pour objectif la modélisation des bâtiments à l'aide de modèles thermiques, un volet social qui cherche à identifier les habitudes et les comportements des habitants via des données d'enquêtes et socio-démographiques, et un volet data qui cherche à modéliser les comportements et leur évolution à partir de données de capteurs.

Afin de mener à bien l'ensemble de ces objectifs, une partie importante de ce projet est la collecte de données au sein de trois bâtiments. Cette instrumentation des bâtiments consiste d'abord à la collecte de données décrivant l'environnement des parties communes des trois bâtiments ainsi que des données de consommation énergétique de la VMC et de l'ascenseur. Afin d'accompagner ces données, une station météo a été installée à proximité du parc de

logements. Ensuite, le projet prévoit la collecte de données de consommation d'électricité à une échelle plus ou moins fine, ainsi que des données d'ambiance et de présence au sein d'un ensemble d'appartements au sein des bâtiments. Pour finir, des enquêtes ont été réalisées pour collecter des données sur la situation socio-démographique du ménage, les habitudes de consommation et d'occupation et les sensibilités vis-à-vis de l'économie d'énergie. La collecte des données au sein des appartements est toujours en cours.

Cette section permet de présenter les données collectées au sein des parties communes des bâtiments. Trois bâtiments de taille variable (2, 5 et 7 étages) ont été instrumentés.

2.4.1 Température intérieure

Afin de collecter des données de température et d'humidité au sein des parties communes, 3 capteurs par bâtiment ont été installés à différents étages. Les capteurs permettent de collecter des données au pas de temps horaire. Les données sont disponibles du 18 février 2019 jusqu'au 25 juin 2020.

Par la suite, et dès que les données ANDRE des parties communes seront évoquées, les capteurs et bâtiments seront désignés tels que présentés dans le tableau ci-dessous :

TABLE 2.2 – Emplacement et appellation des capteurs de température et d'humidité dans les parties communes des bâtiments.

Bâtiment	Étage	Appellation du capteur
Bâtiment 1 (2 étages)	Rez-de-chaussée	B1_E0
	1er étage	B1_E1
	2ème étage	B1_E2
Bâtiment 2 (5 étages)	Rez-de-chaussée	B2_E0
	2ème étage	B2_E2
	4ème étage	B2_E4
Bâtiment 3 (7 étages)	Rez-de-chaussée	B3_E0
	3ème étage	B3_E3
	6ème étage	B3_E6

Les données manquantes sont peu nombreuses et on n'observe pas de longues périodes de données manquantes. Par conséquent, l'interpolation des données ne posera pas de problème par la suite.

Les figures 2.11 et 2.10 représentent les données de température mesurée par les capteurs des parties communes. On peut observer des évolutions au cours de l'année selon la saison et le mois de l'année ou encore selon le jour de la semaine. On observe également des disparités dans les températures mesurées d'un capteur à l'autre. Cela indique que la température à chaque étage et chaque bâtiment n'est pas uniquement déterminée par des variables calendaires ou des données météorologiques communes. On peut également ajouter que les températures mesurées au sein des parties communes sont proches au cours des mois

d'été et de printemps (de mai à septembre) tandis que les courbes sont plus différentes pendant la période de chauffage allant d'octobre à avril. On peut donc imaginer que les parties communes ne sont pas chauffées de la même manière.

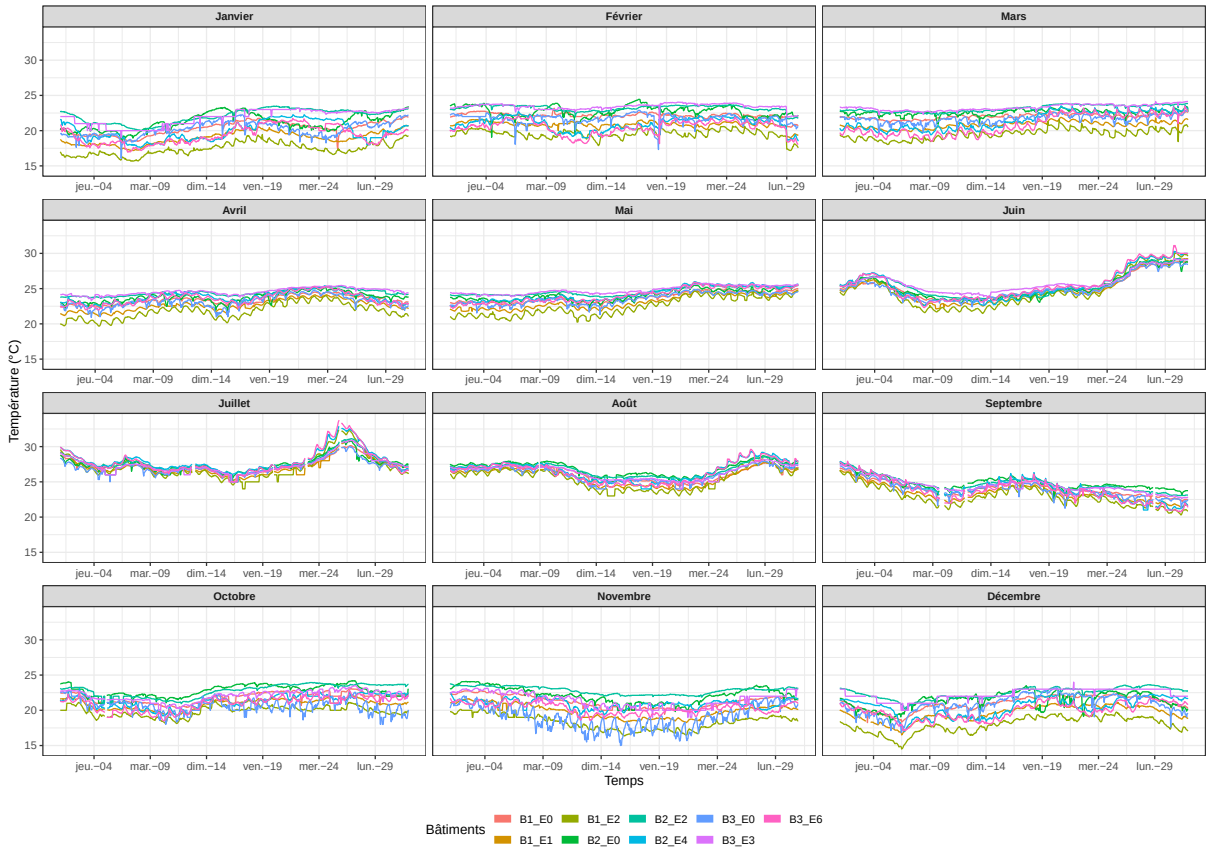

FIGURE 2.10 – Température intérieure au cours de chaque mois de l'année pour les différents capteurs des parties communes de la base ANDRE. Pour les capteurs dont les données sont disponibles sur plusieurs années, les données sont moyennées.

Ces données pourront être utilisées dans le cadre de l'application sur des données réelles des modèles présentés dans le chapitre 4. Ensuite, dans un objectif de classification et d'identification d'effets exogènes relatifs à des facteurs exogènes communs, les données météorologiques peuvent être utiles.

2.4.2 Les données météorologiques

Afin de collecter des données sur les conditions météorologiques au cours de la période de collecte de données au sein des bâtiments, une station météo est installée proche du parc de logements à partir du 2 février 2020. Pour faire correspondre les périodes de collecte des données dans les parties communes et les données météorologiques, un autre jeu de données pourra être utilisé dans ce cas. Il s'agit d'une base de données issue d'une station météo située dans la ville de Lagny-sur-Marne, peu éloignée du parc de logements. Les données sont disponibles du 1er janvier 2018 jusqu'au 30 avril 2020 au pas de temps horaire. Au cours de

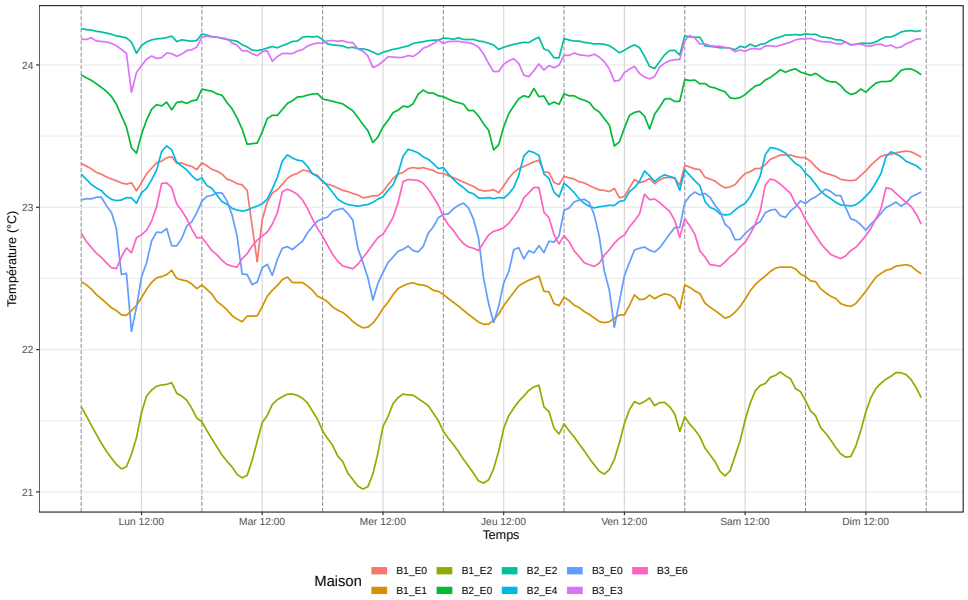

FIGURE 2.11 – Courbe hebdomadaire moyenne de température intérieure pour les différents capteurs des parties communes de la base ANDRE.

cette période, aucune valeur n'est manquante dans la base de données dont nous disposons. Nous n'avons pas d'information sur d'éventuels retraitements réalisés sur les données.

Cette station météo permet de collecter les données de température extérieure (°C), d'humidité (%), les précipitations (mm/h²) et la température ressentie (°C) qui dépend, entre autres, de la vitesse du vent et de la température extérieure. Les données sont présentées dans le graphique 2.12.

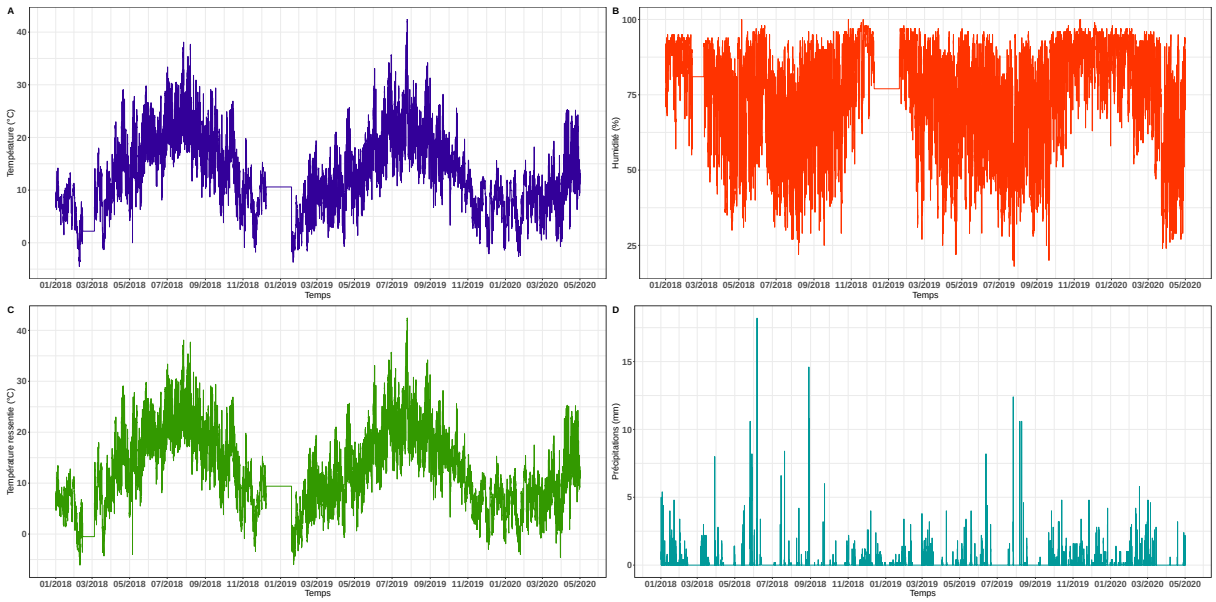

FIGURE 2.12 – Données météorologiques sur l'ensemble de la période d'observation de la base ANDRE au pas de temps horaire. (A) représente la température extérieure, (B) représente l'humidité extérieure, (C) représente l'irradiance solaire et (D) représente les précipitations horaires moyennes.

2.5 Conclusion

Dans le domaine de l'analyse de données énergétiques, l'un des objectifs principaux est la prédiction de consommation d'énergie à l'échelle d'un logement, d'un bâtiment ou d'une ville. Les auteurs s'accordent sur le fait que les comportements des habitants, en matière de chauffage, de présence ou d'habitude de consommation, ont un effet important sur la consommation, mais sont variables et dépendent de nombreux facteurs, ce qui rend l'estimation de ces comportements difficile. Afin d'améliorer la connaissance de ces comportements, résumer l'information, de nombreux travaux portent sur leur classification afin d'identifier des comportements type. L'analyse des clusters permet de les caractériser à partir de variables calendaires ou socio-démographiques sur les ménages. La dynamique des classes et l'évolution des comportements extraits via la classification constituent des sujets importants, intéressants et peu explorés. Par conséquent, l'objectif des travaux qui seront présentés par la suite est de modéliser la dynamique des comportements dans un modèle de classification.

Les données qui seront utilisées pour l'application des modèles sont issues de deux jeux de données. Le premier regroupe des données de maisons individuelles anglaises. La température, l'humidité et le nombre de détections de mouvement au sein des pièces de vie sont disponibles. De plus, des données concernant le contexte météorologique pourront être utilisées. Ce chapitre a également permis de présenter les jeux de données constituées de données de température, d'humidité mesurée au sein de maisons individuelles ou de parties communes d'immeubles. Ces deux jeux de données contiennent également des variables météorologiques et socio-démographiques pour caractériser les maisons.

Chapitre 3

Méthodes et outils

Contents

3.1 Introduction

Ce chapitre permet d'introduire les différentes méthodes, modèles et outils qui seront utilisés par la suite. Dans un premier temps, on aborde les modèles de mélange dédiés à la classification non supervisée de données temporelles. Plus particulièrement, les mélanges gaussiens et les mélanges de régressions seront d'abord présentés, puis un modèle de mélange incorporant un *a priori* sur la dynamique des clusters. Un aspect important de la classification résidant dans l'estimation des paramètres du modèle construit, la seconde section de ce chapitre est destinée à présenter deux méthodes d'estimation adaptées aux modèles de mélange. Dans le cas de modèles à variables latentes où le maximum de vraisemblance n'est pas directement calculable, l'algorithme EM est couramment utilisé [McLachlan and Krishnan, 2007; Dempster et al., 1977a]. Cependant, l'utilisation de cet algorithme requiert une forme littérale de la log-vraisemblance, ce qui n'est pas évident dans le cas de modèles plus complexes. Dans ce cas, les méthodes d'inférence variationnelle et l'algorithme VEM peuvent être utilisés [Blei et al., 2017]. Pour finir, une des problématiques centrales dans la classification étant le choix du nombre de clusters, la dernière section est dédiée aux critères classiques pour la sélection de modèle.

3.2 Les modèles de mélange pour la classification non supervisée de données temporelles

Dans la littérature, de nombreux modèles de classification ont été développés afin de regrouper des données en classes homogènes. Les modèles de mélange constituent une famille importante de modèles de classification probabiliste [McLachlan, 2015]. Dans ce cas, les auteurs supposent que les données à classifier proviennent du mélange d'un nombre fini de densités dans certaines proportions inconnues. En d'autres termes, chaque observation est considérée comme appartenant à une classe à laquelle est associée une densité de probabilité. Par conséquent, les données sont des réalisations d'une variable aléatoire distribuée suivant une combinaison linéaire de densités de probabilité. Les modèles de mélange permettent une grande flexibilité dans le choix des distributions. En effet, il est possible de construire des modèles de mélange gaussiens [Eirola and Lendasse, 2013], qui sont la forme la plus classique pour des données continues, mais on peut également utiliser d'autres distributions comme des lois de Poisson dans le cas de données de comptage [Randriamanamihaga et al., 2014]. Les mélanges de régressions [Desarbo and Cron, 1988; Hurn et al., 2000] permettent d'introduire des composantes régressives afin d'estimer l'effet de variables mesurables et connues au sein de chaque cluster identifié. Dans le cas de données temporelles, il est intéressant de modéliser, sous la forme de processus stochastiques, la dynamique des centres de classes [El Assaad et al., 2016].

Cette section est, dans un premier temps, dédiée aux modèles de mélange gaussiens, qui sont une forme classique et largement utilisée des modèles de mélange. Puis, nous présenterons les mélanges de régressions ainsi que des modèles avec un *a priori* sur la dynamique des centres de classes. L'objectif est de poser les bases des modèles de mélange afin de mieux appréhender les travaux qui seront présentés par la suite.

3.2.1 Les modèles de mélange gaussiens

Les modèles de mélange gaussiens sont couramment utilisés pour la classification de données continues. Ces derniers permettent de modéliser la densité d'un jeu de données comme un mélange de distributions gaussiennes [McLachlan and Krishnan, 2007; McLachlan and Peel, 2004].

Les modélisations présentées par la suite concernent des données qu'on note $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, et $\mathbf{x}_i \in \mathbf{R}^d$. On fait l'hypothèse que ces observations sont indépendantes et générées par un mélange de K densités de probabilité dans des proportions inconnues. Chaque observation appartient à une des K classes qui sont modélisées sous forme d'une variable qu'on notera $\mathbf{z}$. Cette variable modélise la classification cachée $\mathbf{z} = (z_1, ...z_n)$, où $z_i \in \{1, ..., K\}$ indique la classe d'appartenance de la i$^{\text{ème}}$ observation. On note (π_k) les proportions du mélange avec $\pi_k = P(z_i = k)$. La densité de probabilité de l'observation $\mathbf{x}_i$ se définit par :

$$p(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \varphi_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.1}$$

avec $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ les paramètres des distributions normales du mélange.

La figure 3.1(A) représente un exemple de jeu de données simulées à partir d'un mélange gaussien à deux composantes en deux dimensions. De plus, les figures 3.1(B) et 3.1(C) représentent les deux densités des composantes du mélange et la densité jointe dont sont issues les données de la figure 3.1(A).

Étant donné que les n observations sont indépendantes, on peut écrire la log-vraisemblance $L(\mathbf{x}; \boldsymbol{\Theta})$ et la log-vraisemblance complétée $L_c(\mathbf{x}, \mathbf{z}; \boldsymbol{\Theta})$ de la manière suivante :

$$\begin{aligned} L(\mathbf{x}; \boldsymbol{\Theta}) &= \log p(\mathbf{x}; \boldsymbol{\Theta}) \\ &= \log \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{N} \log(\sum_{k=1}^{K} \pi_k \, \varphi_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)), \end{aligned} \tag{3.2}$$

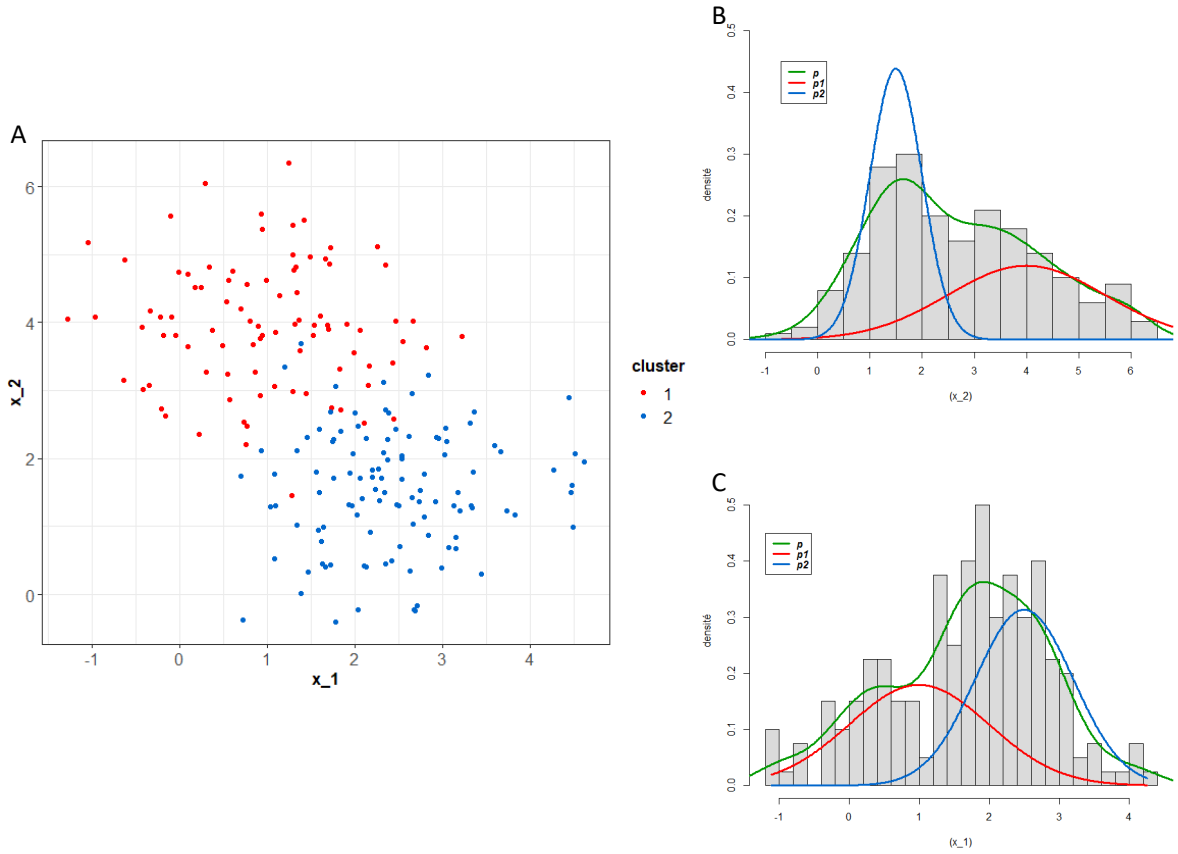

FIGURE 3.1 – Nuage de points et densités d'un jeu de données issues d'un modèle de mélange gaussien à deux composantes avec $K = 2$, en dimension $d = 2$, $n = 200$ observations et dans les proportions $\pi = (0.55, 0.45)$. (A) représente le nuage de points des données $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, et $\mathbf{x}_i \in \mathbb{R}^2$. (B) et (C) représentent les histogrammes du jeu de données généré où p_1 désigne la densité de la première composante avec $\mu_1 = (1, 4)$, $\mathbf{\Sigma}_1 = \text{diag}(1, 1.5)$, p_2 correspond à la densité de la seconde composante avec $\mu_2 = (2.5, 1.5)$, $\mathbf{\Sigma}_2 = \text{diag}(0.7, 0.5)$. La densité p est la densité du jeu de données correspondant au mélange de p_1 et p_2 dans les proportions de mélange π.

et

$$\begin{aligned} L_c(\mathbf{x}, \mathbf{z}; \mathbf{\Theta}) &= \log p(\mathbf{x}, \mathbf{z}; \mathbf{\Theta}) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log(\pi_k \, \varphi_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)), \end{aligned} \tag{3.3}$$

avec $z_{ik} = 1$ si l'observation i appartient à la classe k et 0 sinon. De plus, comme nous sommes dans le cas gaussien, φ_k est la densité d'une loi normale :

$$\varphi_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\mathbf{\Sigma}_k)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)' \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \tag{3.4}$$

Les paramètres du modèle à estimer sont $\mathbf{\Theta} = (\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \boldsymbol{\pi}_k)_{k=1,...K}$. Pour réduire le nombre de paramètres et simplifier le modèle, des contraintes peuvent être ajoutées sur les différents paramètres, selon l'information disponible sur les données. Dans [Celeux and Govaert,

1995], les auteurs proposent différentes variantes du modèle selon la forme des matrices de variances-covariances ($\mathbf{\Sigma}_k$). Les auteurs distinguent quatorze modèles. La volonté de construire des modèles parcimonieux peut amener à contraindre les matrices de variances afin de limiter le nombre de paramètres. Dans [Banfield and Raftery, 1993], les auteurs discutent également le choix des contraintes sur les matrices de variances-covariances. Il est possible de contraindre les matrices de variances en les considérant diagonales et identiques pour les différentes dimensions. On parle alors de cas "sphérique" (cf 3.2(A)). Le cas présenté dans la figure 3.2(B) correspond à un cas où les variables observées sont indépendantes, mais de variances différentes, et les matrices de variances-covariances ne sont pas les mêmes pour les deux classes. Ensuite, la figure 3.2(C) représente un cas avec des matrices de variances-covariances qui ne sont pas diagonales, mais identiques pour les deux classes. Dans le cas où les matrices ne sont pas diagonales, les covariances ne sont pas nulles et on considère que les deux variables ne sont pas indépendantes. Pour finir, la figure 3.2(D) représente le cas non contraint pour lequel les matrices ne sont pas diagonales et diffèrent d'une classe à l'autre.

Le cas contraint de la classification par l'algorithme des K-means

Les modèles de mélange gaussiens sont présentés comme une famille importante et souvent utilisée pour faire de la classification. L'algorithme des K-means est également souvent décrit comme une méthode standard de la classification et sera évoqué dans cette thèse.

La modélisation sous-jacente à l'algorithme des K-means est un modèle de mélange gaussien contraint. Dans [Lücke and Forster, 2019], les auteurs montrent que la modélisation d'une classification K-means peut-être rapportée à un modèle de mélange gaussien avec les contraintes d'égalité des proportions du mélange et des matrices de variances-covariances identiques, diagonales et sphériques.

Soit le jeu de données $\mathbf{x}$ présenté précédemment. L'algorithme de classification des K-means a pour objectif la minimisation du critère suivant :

$$W(\mathbf{z}, \boldsymbol{\mu}, \mathbf{x}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} ||\mathbf{x}_i - \boldsymbol{\mu}_k||^2, \tag{3.5}$$

avec $z_{ik} = 1$ si l'observation i appartient au cluster k et $\boldsymbol{\mu}_k$ les centres de clusters. Afin de minimiser ce critère, l'algorithme K-means consiste, dans un premier temps, à définir des centres de classes pour ensuite attribuer le cluster le plus proche à chaque observation selon la distance Euclidienne. Une fois cette première étape d'attribution réalisée, les centres de classes sont recalculés en moyennant les observations selon leur cluster d'appartenance. Ces deux étapes d'attribution et de recalcul des centres de classes, sont réitérées jusqu'à atteindre une solution stable.

Dans le cas d'un mélange gaussien, on cherche à optimiser la vraisemblance complétée (3.3). Or, si l'on impose les contraintes suivantes :

- $\forall k, \pi_k = 1/K$,

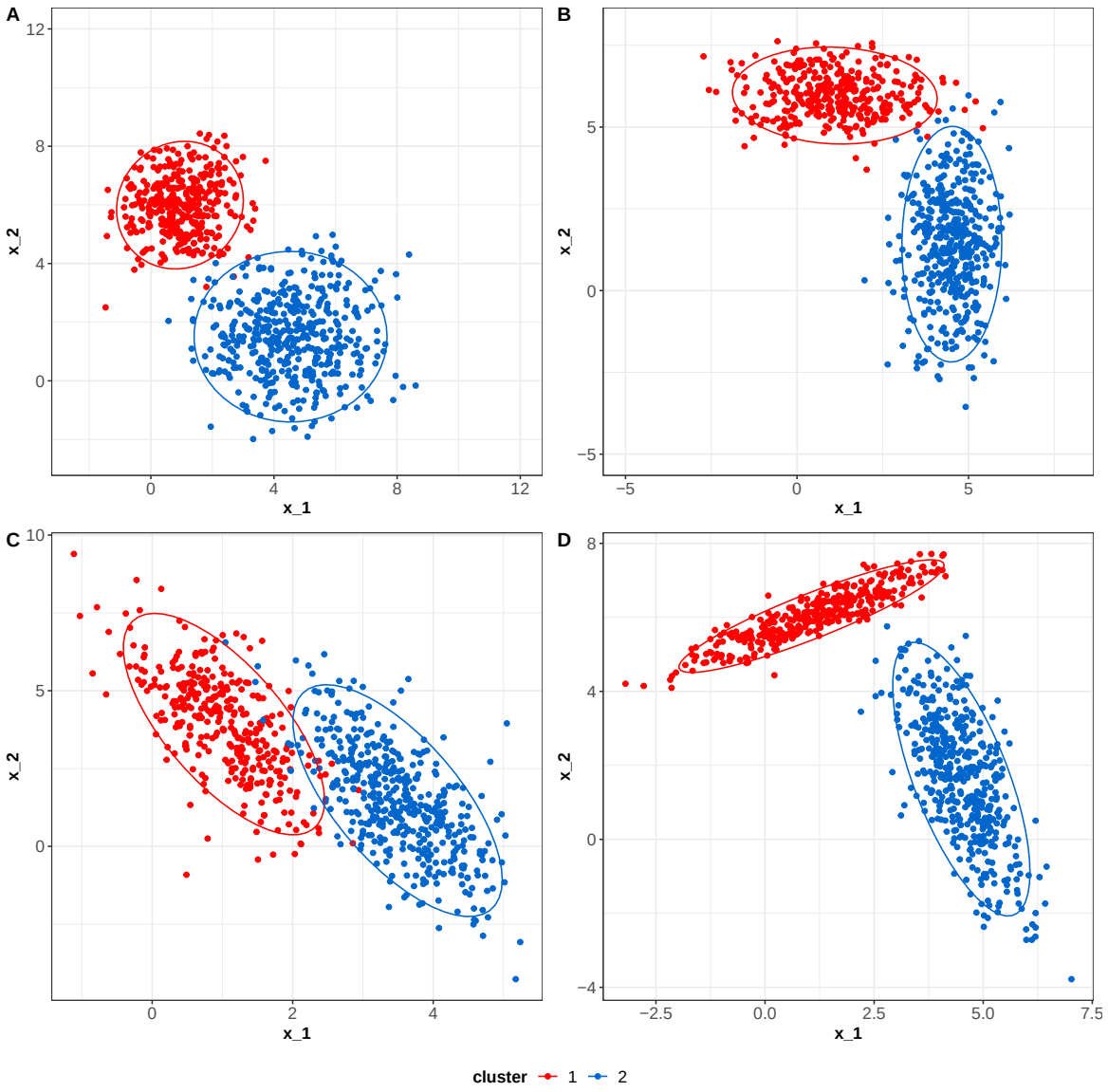

FIGURE 3.2 – Exemple de nuage de points simulés à partir de mélanges gaussiens avec différentes contraintes sur les matrices de variances-covariances avec 800 observations ($n = 800$) et les proportions de mélanges $\boldsymbol{\pi} = (0.56, 0.44)$. (A) représente le sphérique où les covariances sont nulles, et les variances identiques pour les deux dimensions : $\boldsymbol{\Sigma_k} = \sigma_k^2 \mathbb{I}_2$. (B) représente le cas diagonal, où les covariances sont nulles, les variances sont différentes pour les deux variables : $\boldsymbol{\Sigma_k} = \text{diag}(\sigma_{1k}^2, \sigma_{2k}^2)$. (C) représente le cas où les covariances ne sont pas nulles, mais les matrices sont identiques pour les deux classes : $\boldsymbol{\Sigma_k} = \boldsymbol{\Sigma}$. (D) représente le cas le moins contraint avec aucune contrainte sur les matrices de variances-covariances $\boldsymbol{\Sigma_k}$.

- $\forall k, \boldsymbol{\Sigma}_k = \sigma^2 \mathbb{I}_d$,

alors, il est possible de réécrire la vraisemblance complétée, notée $L_c^{KM}(.)$ telle que :

$$\begin{aligned} L_c^{KM}(\mathbf{x}, \mathbf{z}; \boldsymbol{\Theta}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\frac{1}{K} \varphi_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \sigma^2 \mathbb{I}_d) \\ &\propto -\frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{k=1}^{K} (z_{ik} ||\mathbf{x}_i - \boldsymbol{\mu}_k||^2) - nd \log(2\pi\sigma^2). \end{aligned} \tag{3.6}$$

Dans ce cas, la maximisation de la vraisemblance (3.6) par rapport à $\mathbf{z}$ et $\boldsymbol{\mu}$, revient à la minimisation du critère (3.5).

3.2.2 Les modèles de mélange de régressions

La modélisation via des modèles de mélange sous-entend qu'il existe des classes inobservées et que les données sont issues d'un mélange de plusieurs distributions. Dans le cas où il existe des variables exogènes connues, les mélange de régressions permettent de classifier les observations et d'estimer des effets régressifs propres à chaque classe [Desarbo and Wedel, 2002; Hurn et al., 2000]. Les modèles de mélange de régressions permettent de modéliser des données comme un mélange de lois dont les paramètres dépendent d'un ensemble de variables mesurables. L'objectif est de classifier les observations en estimant des coefficients et des matrices de variances-covariances propres à chaque cluster. La famille des modèles de mélange de régressions regroupe plusieurs modèles et variantes, mais cette sous-section a pour objectif de faire une présentation générale des deux cas possibles d'application des modèles de mélange de régressions destinés aux données temporelles.

Soient les données $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, et $\mathbf{x}_i = (x_{i1}, ..., x_{iT}) \;\; x_{it} \in \mathbf{R}$. On considère dans ce cas que les observations sont issues de K distributions. Mais on fait également l'hypothèse que les données temporelles sont soumises à des effets de variables temporelles exogènes, communes à l'ensemble des observations $(\mathbf{x}_i)$, $\mathbf{u} = (\mathbf{u}_1, ..., \mathbf{u}_T)$ et $\mathbf{u}_t = (u_{t1}, ..., u_{tp})$. Dans le cas de la régression polynomiale, on a par exemple $\mathbf{u}_t = (1, t, t^2, \ldots)$.

Il est important pour la suite de distinguer deux cas de figure dans les modèles de classification sur des données temporelles. En effet, il est possible de construire un modèle de mélange de régressions pour la classification des observations x_{it} et ainsi considérer que chaque observation peut appartenir à une classe différente selon le pas de temps. La seconde possibilité est de construire un modèle de mélange de régressions pour la classification des séquences $\mathbf{x}_i$. Dans ce cas, l'observation i appartient à la même classe pour toute la séquence $(x_{i1}, \ldots, x_{iT})$. Les deux modèles vont être présentés par la suite afin de bien distinguer les deux cas de figure et de positionner les travaux qui seront présentés par la suite.

Classification des observations (x_{it})

On considère les données $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, et $\mathbf{x}_i = (x_{i1}, ..., x_{iT}) \;\; x_{it} \in \mathbf{R}$. Comme dit précédemment, on suppose que les observations (x_{it}) sont générés indépendamment par K distributions et sont sujettes à des effets régressifs d'un ensemble de p variables exogènes $\mathbf{u}_t = (u_{t1}, ..., u_{tp})$. Dans ce cas, si z_{it} désigne la variable latente liée à la classe d'appartenance de l'observation i au temps t, le mélange de régressions peut être défini comme suit :

$$x_{it} = \boldsymbol{\beta}'_{z_{it}} \mathbf{u}_t + \epsilon_{it}, \;\; \text{avec} \;\; \epsilon_{it} \sim \mathcal{N}(0, \sigma^2_{z_{it}}). \tag{3.7}$$

Ces éléments permettent de définir la loi de probabilité de x_{it} comme suit :

$$p(x_{it};\boldsymbol{\beta},\boldsymbol{\sigma^2},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \varphi(x_{it};\boldsymbol{\beta}_k'\mathbf{u}_t,\sigma_k^2), \tag{3.8}$$

avec $(\boldsymbol{\beta}_k)$ le vecteur de coefficients associé aux variables $\mathbf{u}_t$, σ_k^2 les variances et π_k les proportions du mélange. On note $\boldsymbol{\Theta}$ le vecteur de paramètre du modèle. Étant donné que les n séquences sont indépendantes, on peut écrire la log-vraisemblance de $\mathbf{x}$ et la log-vraisemblance complétée de $(\mathbf{x},\mathbf{z})$ sous la forme suivante :

$$L(\mathbf{x};\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{t=1}^{T} \log(\sum_{k=1}^{K} \pi_k\, \varphi_k(x_{it};\boldsymbol{\beta}_k'\mathbf{u}_t,\sigma_k^2)), \tag{3.9}$$

$$L_c(\mathbf{x},\mathbf{z};\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ikt}\, \log(\pi_k\, \varphi_k(x_{it};\boldsymbol{\beta}_k'\mathbf{u}_t,\sigma_k^2)), \tag{3.10}$$

avec $z_{ikt} = 1$ si l'observation x_{it} appartient à la classe k et 0 sinon et φ_k désigne la densité d'une loi normale.

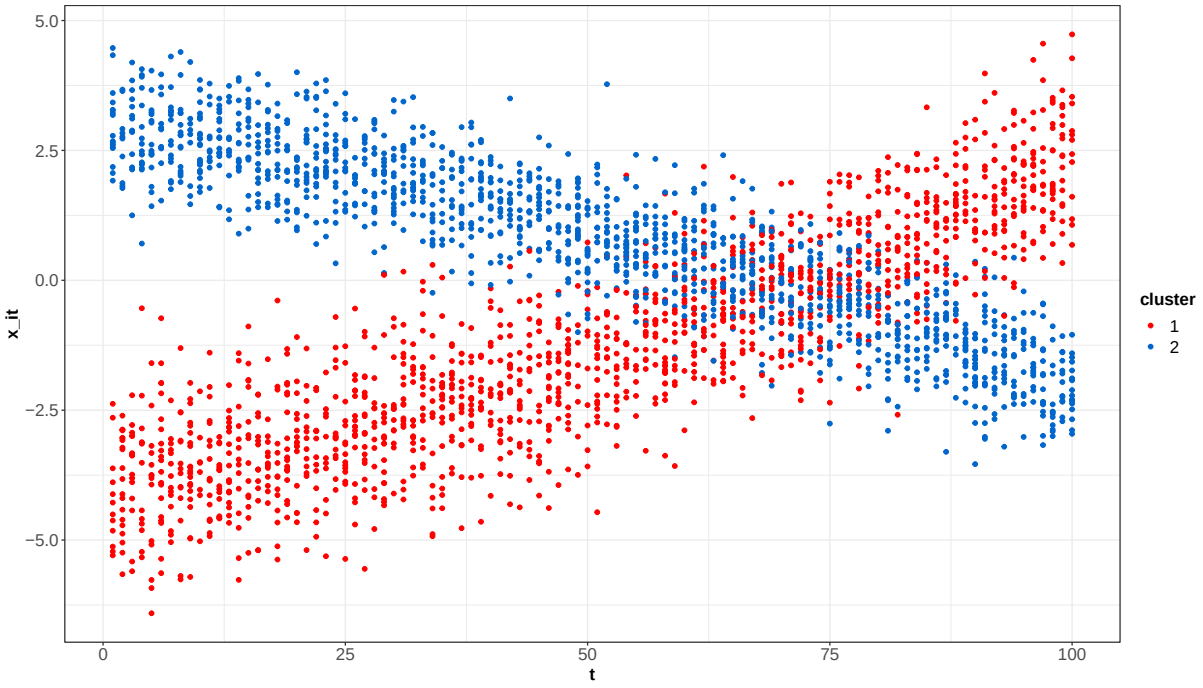

FIGURE 3.3 – Exemple de jeux de données simulées à partir d'un mélange de régressions où chaque observation x_{it} appartient à une classe. Les données sont générées avec deux composantes ($K = 2$), deux facteurs exogènes ($p = 2$). Le jeu de données contient $n = 30$ observations et $T = 100$ séquences de temps. Les proportions de mélange sont $\boldsymbol{\pi} = (0.5, 0.5)$.

La figure 3.3 représente des données (x_{it}) générées à partir d'un mélange de régressions à deux classes. Pour rappel, dans ce cas, chaque observation $\mathbf{x}_{it}$ appartient à un cluster qui peut varier au cours du temps.

Classification des séquences $(\mathbf{x}_i)$

On fait l'hypothèse que les observations $\mathbf{x}_i = (x_{i1}, ..., x_{iT})$ sont indépendantes, issues de K distributions et sujettes à des effets régressifs d'un ensemble de p variables exogènes $\mathbf{u}_t = (u_{t1}, ..., u_{tp})$. On considère les séquences d'observation $\mathbf{x}_i$ avec z_i la variable latente désignant la classe d'appartenance de l'observation i. Le modèle est le suivant :

$$x_{it} = \boldsymbol{\beta}'_{z_i}\mathbf{u}_t + \epsilon_{it}, \text{ avec } \epsilon_{it} \sim \mathcal{N}(0, \sigma^2_{z_i}). \tag{3.11}$$

Ces éléments permettent de définir la loi de probabilité de la séquence $\mathbf{x}_i$ comme suit :

$$p(\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} \varphi(x_{it}; \boldsymbol{\beta}'_{z_i}\mathbf{u}_t, \sigma^2_k), \tag{3.12}$$

avec $(\boldsymbol{\beta}_k)$ le vecteur de coefficients associés aux variables $\mathbf{u}_t$, σ^2_k les paramètres de variances et π_k les proportions du mélange. Étant donné que les observations sont indépendantes, on peut écrire la log-vraisemblance $L(\mathbf{x}; \boldsymbol{\Theta})$ et la log-vraisemblance complétée de $(\mathbf{x}, \mathbf{z})$ telles que :

$$\begin{aligned} L(\mathbf{x}; \boldsymbol{\Theta}) &= \log p(\mathbf{x}; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} \varphi_k(x_{it}; \boldsymbol{\beta}'_k\mathbf{u}_t, \sigma^2_k)\Big), \end{aligned} \tag{3.13}$$

$$\begin{aligned} L_c(\mathbf{x}, \mathbf{z}; \boldsymbol{\Theta}) &= \log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\Theta}) = \log p(\mathbf{x}|\mathbf{z}; \boldsymbol{\Theta})\, p(\mathbf{z}; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log\Big(\pi_k \prod_{t=1}^{T} \varphi_k(x_{it}; \boldsymbol{\beta}'_k\mathbf{u}_t, \sigma^2_k)\Big), \end{aligned} \tag{3.14}$$

avec $z_{ik} = 1$ si la séquence $\mathbf{x}_i$ appartient à la classe k et 0 sinon et φ_k désigne la densité d'une loi normale. La figure 3.4 représente des données (x_{it}) générées à partir d'un mélange de régressions à deux classes pour la classification des séquences $\mathbf{x}_i$. À noter que la séquence $\mathbf{x}_i$ appartient à un seul cluster pour l'ensemble des pas de temps t observés.

L'avantage, comme pour les mélanges gaussiens, des mélanges de régressions est leur grande flexibilité. En effet, il existe de nombreuses variantes des modèles présentés ci-dessus. Parmi ces variantes, qui semblent pertinentes dans le cadre de données temporelles, on peut citer les travaux de [Devijver et al., 2015]. Dans cet article, les auteurs proposent une décomposition en ondelettes des séries temporelles et modélisent ensuite les données comme un mélange de régressions avec comme variables régressives les composants identifiés lors de la phase de décomposition. L'avantage de ces travaux est qu'ils proposent un modèle de classification de données temporelles en très grande dimension en identifiant les composants ayant des effets régressifs les plus importants. Il est également éclairant de citer ici les travaux de [Wang et al., 2015] qui proposent de modéliser des données temporelles via un mélange

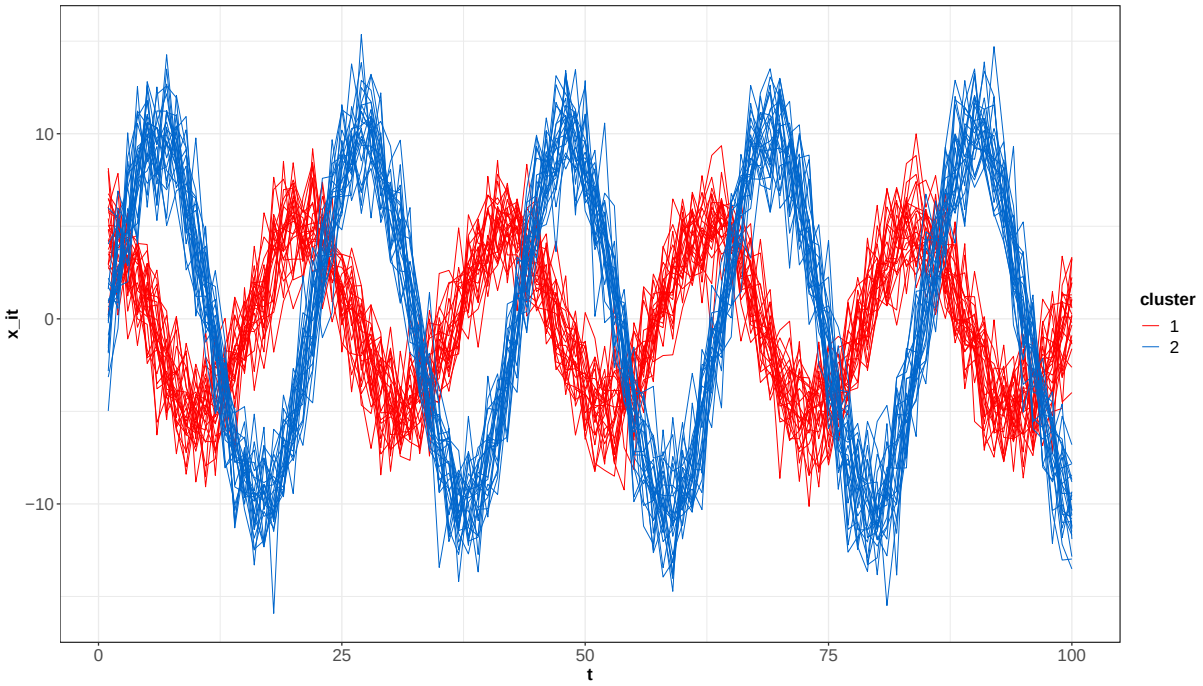

FIGURE 3.4 – Exemple de jeux de données simulées à partir d'un mélange de régressions où la séquence x_i appartient à une classe. Les données sont générées avec deux composantes ($K = 2$), deux facteurs exogènes ($p = 2$). Le jeu de données contient $n = 30$ observations et $T = 100$ séquences de temps. Les proportions de mélange sont $\boldsymbol{\pi} = (0.5, 0.5)$.

de régressions polynomiales. Ces travaux illustrent la variété des mélanges de régressions possibles. En effet, les deux cas généraux présentés plus haut sont des cas de régression linéaire simples, mais il est possible d'imaginer que les données soient soumises à des effets non-linéaires.

Dans [Jones and McLachlan, 2008] ou [Basford and McLachlan, 1985], les auteurs appliquent des modèles de régressions à des données en trois dimensions et les centres de classes sont donc également considérés comme tridimensionnels. Cela permet d'utiliser ces modèles dans des domaines variés comme la classification de données de génotypes.

3.2.3 Modèles de mélange à centres de classes dynamiques

Nous avons vu comment modéliser des données comme un mélange gaussien, ou comme un mélange de régressions, dans le cadre de données temporelles. Il est également possible de modéliser des données temporelles sous la forme d'un mélange de densités gaussiennes en ajoutant une loi *a priori* sur la dynamique temporelle des centres de classes. Les modèles de mélange gaussiens classiques, même appliqués à des données temporelles, ne permettent pas de modéliser la dépendance temporelle des observations et la dynamique des centres de classes. Les modèles de régressions sont de meilleurs candidats pour prendre en compte des effets dépendant du temps. À cette fin, ces modèles nécessitent de prédéfinir les facteurs temporels. Afin de modéliser ces dynamiques sans avoir à définir des facteurs exogènes,

plusieurs auteurs ont développé des modèles qui permettent de modéliser la dynamique des centres de classes en fixant un *a priori* sur ces centres. Parmi ces travaux, on peut citer les modèles de mélange présentés dans [Calabrese and Paninski, 2011] ou dans [El Assaad et al., 2016] qui constituent une bonne introduction à ces modèles dynamiques. Dans ces deux articles, les auteurs ont développé des modèles de mélange pour des données temporelles en considérant que les centres de classes sont des processus stochastiques. Nous présenterons le modèle de classification à centres de classes dynamiques développé dans [El Assaad et al., 2016].

MODÈLES DE MÉLANGE À CENTRES DYNAMIQUES ET CLASSIFICATION DE DONNÉES TEMPORELLES

Dans [El Assaad et al., 2016] les auteurs proposent un modèle de classification de données temporelles avec estimation de la dynamique des centres classes. L'intérêt est double : classifier des données temporelles et les modéliser comme un mélange de lois normales dont les centres, notés $\boldsymbol{\mu}_{kt}\ \forall k,t$, sont temporellement dépendants et sont considérés comme des processus stochastiques. Ce modèle suppose que chaque observation $\mathbf{x}_{it}$ appartient à un cluster k. Une séquence $\mathbf{x}_i$ peut donc appartenir à différents clusters selon l'instant t.

Considérant des données temporelles $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, et $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})\ \ \mathbf{x}_{it} \in \mathbf{R}^d$, le modèle proposé est le suivant :

$$\begin{cases} \mathbf{x}_{it} = \sum_{k=1}^{K} z_{tik}(\boldsymbol{\mu}_{kt} + \sigma_k \boldsymbol{\epsilon}_{it}) \\ \boldsymbol{\mu}_{kt} = \boldsymbol{\mu}_{kt-1} + \nu_k \boldsymbol{\eta}_t, \end{cases} \tag{3.15}$$

avec $z_{tik} = 1$ si l'observation $\mathbf{x}_{it}$ appartient à la classe k. Les variables $(\boldsymbol{\eta}_k)$ et $(\boldsymbol{\epsilon}_{it})$ sont des bruits gaussiens centrés et réduits et ν_k l'écart-type associé au bruit $\boldsymbol{\eta}_k$. Les variables $(\mathbf{z}, \boldsymbol{\mu})$ sont les variables latentes de ce modèle. Le graphique 3.5 représente un jeu de données généré à partir de ce modèle avec quatre classes en dimension une.

On peut écrire la log-vraisemblance du modèle :

$$L(\mathbf{x};\boldsymbol{\Theta}) = \log p(\mathbf{x};\boldsymbol{\Theta}) = \log \sum_{z} p(\mathbf{z};\boldsymbol{\Theta}) \int_{\boldsymbol{\mu}} p(\boldsymbol{\mu}|\mathbf{z};\boldsymbol{\Theta})\, p(\mathbf{x}|\boldsymbol{\mu},\mathbf{z};\boldsymbol{\Theta})\, d\boldsymbol{\mu}, \tag{3.16}$$

avec $\boldsymbol{\Theta} = \{(\pi_k, \boldsymbol{\mu}_{k0}, \sigma_k^2, \nu_k)_{k=1,...,K}\}$, le vecteur de paramètres du modèle. Cette log-vraisemblance n'est pas calculable dans la mesure où il faudrait pouvoir calculer la somme d'intégrales en fonction de l'ensemble des valeurs possibles des variables latentes $(\mathbf{z}, \boldsymbol{\mu})$. La méthode d'estimation utilisée est l'inférence variationnelle et sera présentée dans la section dédiée (3.3).

Le modèle présenté ci-dessus conduit à estimer une classe d'appartenance à chaque pas de temps. De ce fait, une série temporelle appartient à différentes classes, selon l'instant. On peut également construire un modèle de classification avec une modélisation dynamique des centres de classes pour lequel on considère que la classe d'appartenance est unique pour l'ensemble de la séquence [Devijver et al., 2015]. Le modèle présenté dans le chapitre 4 se

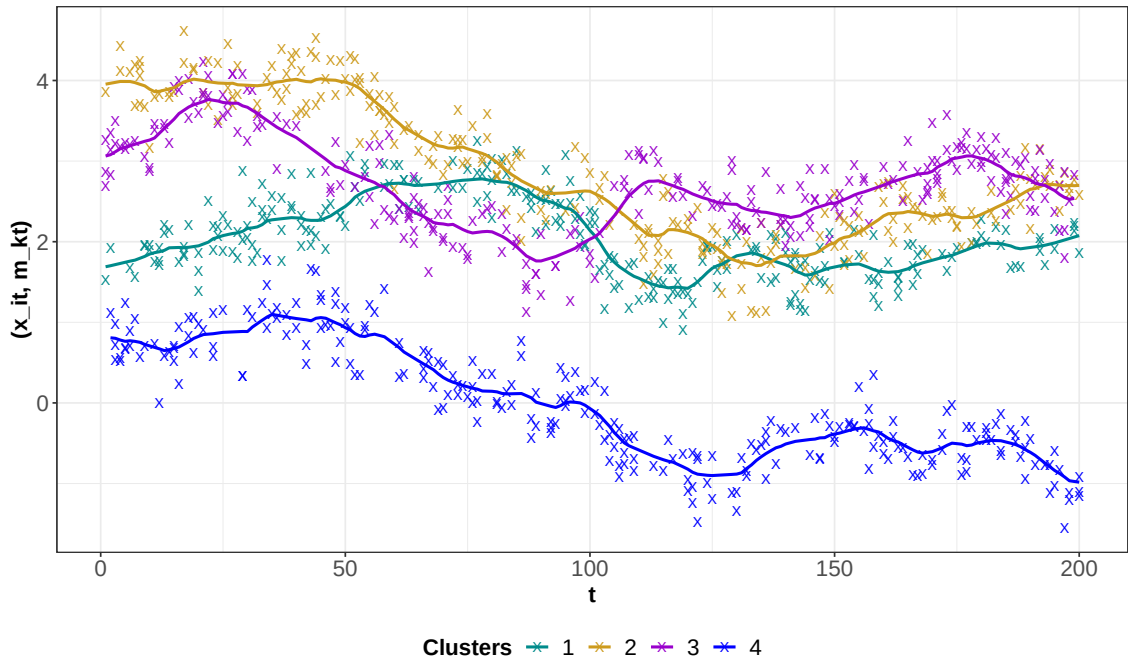

FIGURE 3.5 – Exemple d'un jeu de données généré à partir du modèle de classification dynamique avec centres dynamiques. Les données sont générées avec $K = 4$, $d = 1$, $T = 200$ et $n = 5$. Les proportions de classes sont $\pi = (0.25, 0.25, 0.25, 0.25)$ avec les paramètres de variance $\sigma_k = \frac{1}{16}, \forall k$, et les paramètres de départ $\boldsymbol{\mu}_0 = (2, 4, 3, 1)$. Les observations $\mathbf{x}_{it}$ sont représentées par des croix et les centres de classes par les lignes.

base sur cette idée dans le but de classifier des données tout en estimant les centres de classes sous la forme de processus stochastiques.

Pour finir, d'autres auteurs ont développé des modèles de mélange dynamiques. Par exemple, [Xia and Tang, 2019] proposent un modèle de mélange de chaînes de Markov cachées. Ce modèle permet de capter l'hétérogénéité d'une classe à l'autre et au cours du temps. Il s'agit d'une modélisation intéressante de données temporelles, car, en plus d'identifier la classe des observations, elle modélise des états latents pour expliquer les variations et dynamiques des centres de clusters à partir de modèles de Markov cachés.

Cette section a permis d'introduire la notion de variables latentes et les modèles de mélange appliqués à des données temporelles. Partant du modèle le plus simple qui consiste en un mélange de distributions gaussiennes jusqu'à des modèles plus complexes comme le modèle de classification avec centres de classes stochastiques, nous n'avons pas évoqué jusqu'alors les méthodes d'estimation pour ces modèles. La section suivante introduit les algorithmes Expectation-Maximization et les méthodes d'inférence variationnelle qui sont adaptées pour estimer les modèles qui viennent d'être présentés.

3.3 Les méthodes d'estimation

La classification d'observations via les modèles de mélange sous-entend qu'il existe K distribution dont sont issues les données. Chaque observation appartient à une classe inobservée associée à ces distributions. Pour identifier ces classes d'appartenance, il faut estimer les paramètres du modèle. Ces paramètres, notés $\boldsymbol{\Theta}$, sont composés des paramètres de distributions ($(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ dans le cas gaussien par exemple) et des proportions de classes, notées (π_k). Pour estimer ces paramètres, le maximum de vraisemblance est une méthode fréquemment utilisée. L'estimateur du maximum de vraisemblance est consistant et asymptotiquement efficace [Wasserman, 2010]. Cependant, dans le cas des modèles de mélange présentés dans la section 3.2, le maximum de vraisemblance n'est pas toujours calculable de manière analytique. Une solution consiste à approximer ce maximum via des algorithmes itératifs. L'algorithme *Expectation-Maximisation* (EM), présenté en détail dans l'ouvrage [McLachlan and Krishnan, 2007] permet d'estimer les paramètres en approchant le maximum de vraisemblance. Dans le cas de modèles plus complexes, il est difficile d'obtenir une expression littérale de la vraisemblance, ce qui rend l'approche du maximum de vraisemblance encore plus difficile. Par conséquent, une solution possible est de contourner ce problème en bornant la vraisemblance et approchant, itérativement, cette borne. Les méthodes d'inférence variationnelles, décrites dans [Blei et al., 2017] notamment, permettent cela.

3.3.1 L'algorithme EM pour les modèles de mélange

Pour estimer les paramètres d'un modèle de mélange, la méthode du maximum de vraisemblance est souvent utilisée car les estimateurs possèdent des propriétés asymptotiques intéressantes. Dans le cas des modèles de mélange, on cherche à estimer les paramètres des distributions gaussiennes du modèle, mais également à classifier les observations. La vraisemblance à maximiser est la vraisemblance $L(\mathbf{x};\boldsymbol{\Theta})$. Cependant, l'existence de variables cachées relatives à la classification rend la maximisation de la vraisemblance impossible. Dans ce cas, l'algorithme EM permet d'approcher, itérativement, ce maximum de vraisemblance via l'optimisation de l'espérance conditionnelle de la vraisemblance complétée $L_c(\mathbf{x},\mathbf{z};\boldsymbol{\Theta})$.

On considère les données $\mathbf{x}$ présentées plus haut, la vraisemblance complétée d'un modèle de mélange dans le cas général s'écrit :

$$L_c(\mathbf{x},\mathbf{z};\boldsymbol{\Theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\,\log(\pi_k\,\varphi_k(\mathbf{x}_i;\boldsymbol{\theta}_k)), \tag{3.17}$$

avec $(\boldsymbol{\theta}_k)$ les paramètres des distributions $\varphi_k(.)$ du mélange et $\boldsymbol{\Theta} = (\boldsymbol{\theta}_k, \pi_k)_{k=1,..K}$.

L'ALGORITHME EM :

L'algorithme *Expectation Maximisation*, introduit par [Dempster et al., 1977b], est un algorithme itératif. Partant d'une valeur initiale $\boldsymbol{\Theta}^{(0)}$, chaque itération $(q+1)$ consiste à la

mise à jour des paramètres précédents $\boldsymbol{\Theta}^{(q)}$ en maximisant l'espérance conditionnelle de la log-vraisemblance complétée notée $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)})$ définie par :

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) &= \mathbb{E}(L_c(\mathbf{x}, \mathbf{z}; \boldsymbol{\Theta}) \mid \mathbf{x}, \boldsymbol{\Theta}^{(q)}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} p(z_i = k|\mathbf{x}_i; \boldsymbol{\Theta}^{(q)}) \log(\pi_k^{(q)} \varphi_k(x_{it}; \boldsymbol{\theta}_k^{(q)})) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log(\pi_k^{(q)} \varphi_k(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})). \end{aligned} \tag{3.18}$$

Cette expression permet d'introduire les probabilités *a posteriori* :

$$\tau_{ik}^{(q)} = p(z_i = k|\mathbf{x}_i; \boldsymbol{\Theta}^{(q)}) = \frac{\sum_{t=1}^{T} \pi_k^{(q)} \varphi_k(\mathbf{x}_i, \boldsymbol{\theta}_k^{(q)})}{\sum_{t=1}^{T} \sum_{j=1}^{K} \pi_j^{(q)} \varphi_j(\mathbf{x}_i, \boldsymbol{\theta}_j^{(q)})}, \tag{3.19}$$

qui permettent, une fois que l'algorithme a convergé et que l'estimation des paramètres est réalisée, de déterminer la classe d'appartenance *a posteriori* des observations via la méthode du *Maximum a posteriori* (MAP) qui consiste à attribuer la classe correspondant à la probabilité d'appartenance *a posteriori* (τ_{ik}) la plus élevée.

L'algorithme itératif consiste en la répétition des deux étapes suivantes :

Expectation Calculer les probabilités *a posteriori* $\tau_{ik}^{(q+1)}, \forall i = 1,..,n \; k = 1,..K$, à partir de la formule (3.19).

Maximisation Mettre à jour des paramètres $\boldsymbol{\Theta}^{(q+1)} = (\boldsymbol{\theta}_k^{(q+1)}, \pi_k^{(q+1)})_{k=1,..K}$ tel que :

$$\begin{aligned} \boldsymbol{\Theta}^{(q+1)} &= \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(q)}) \\ &= \underset{\boldsymbol{\theta}_k, \pi_k}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q+1)} \log(\pi_k \varphi_k(\mathbf{x}_i; \boldsymbol{\theta}_k)). \end{aligned} \tag{3.20}$$

L'algorithme permet de mettre à jour, à chaque itération, les paramètres du modèle et les probabilités *a posteriori*.

Il existe un grand nombre de variantes de cet algorithme en fonction du modèle considéré ou des paramètres d'intérêt.

Classification Expectation Maximization

Parmi ces variantes, on peut citer l'algorithme **CEM**, proposé par [Celeux and Govaert, 1992]. Cette variante consiste, une fois les probabilités *a posteriori* mises à jour au cours de l'étape "E", à classifier les observations à l'aide de la règle du MAP, pour ensuite résoudre le problème de maximisation afin d'estimer les paramètres. L'algorithme **CEM** se base sur la maximisation de la log-vraisemblance complétée et consiste en la répétition des trois étapes suivantes :

Expectation Calculer les probabilités *a posteriori* $\tau_{ik}^{(q+1)}, \forall i = 1,..,n \; k = 1,..K$, à partir de la

formule (3.19).

Classification Estimer la classe d'appartenance $(z_i^{(q+1)})$:

$$z_i^{(q+1)} = \underset{k}{\operatorname{argmax}}\, \tau_{ik}^{(q+1)}. \tag{3.21}$$

Maximisation Mettre à jour les paramètres $\boldsymbol{\Theta}^{(q+1)} = (\boldsymbol{\theta}_k^{(q+1)}, \pi_k^{(q+1)})_{k=1,..K}$ tel que :

$$\boldsymbol{\Theta}^{(q+1)} = \underset{\boldsymbol{\theta}_k, \pi_k}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(q+1)} \log(\pi_k\, \varphi_k(\mathbf{x}_i; \boldsymbol{\theta}_k)), \tag{3.22}$$

avec $z_{ik} = 1$ si $z_i = k$ et $z_{ik} = 0$ sinon.

STOCHASTIC EXPECTATION MAXIMISATION (SEM)

Dans l'article [Celeux and Govaert, 1992], les auteurs proposent également une version stochastique appelée **Stochastic Expectation Maximization** (SEM). Cet algorithme a pour particularité d'attribuer une classe d'appartenance aux observations après l'étape "E" en utilisant un tirage aléatoire à partir des probabilités *a posteriori* mises à jour. Ensuite, lors de l'étape de maximisation, les paramètres sont estimés à partir de la partition tirée à l'étape précédente. L'algorithme **SEM** consiste en la répétition des 3 étapes suivantes :

Expectation Calculer les probabilités *a posteriori* $\tau_{ik}^{(q+1)}, \forall\, i = 1,..,n\ k = 1,..K$, à partir de la formule (3.19).

Tirage de classe d'appartenance Tirage aléatoire de la classe d'appartenance $(z_i^{(q+1)})$ selon la loi :

$$z_i^{(q+1)} \sim \mathscr{M}(\tau_{ik}^{(q+1)}), \tag{3.23}$$

où $\mathscr{M}$ désigne la loi multinomiale.

Maximisation Mettre à jour les paramètres $\boldsymbol{\Theta}^{(q+1)} = (\boldsymbol{\theta}_k^{(q+1)}, \pi_k^{(q+1)})_{k=1,..K}$ tel que :

$$\boldsymbol{\Theta}^{(q+1)} = \underset{\boldsymbol{\theta}_k, \pi_k}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(q+1)} \log(\pi_k\, \varphi_k(\mathbf{x}_i; \boldsymbol{\theta}_k)), \tag{3.24}$$

avec $z_{ik} = 1$ si $z_i = k$ et $z_{ik} = 0$ sinon.

Plusieurs travaux ont été réalisés afin de comparer les performances de ces différents algorithmes. Le lecteur peut se référer aux œuvres de [McLachlan and Krishnan, 2007], [Biernacki et al., 2003] ou encore [Celeux and Govaert, 1992].

L'algorithme EM est utilisé pour estimer les paramètres des modèles de mélange de distribution, y compris les mélanges de régressions. Cependant, dans certains cas, les méthodes d'inférence variationnelle sont plus appropriées.

3.3.2 L'inférence variationnelle et l'algorithme VEM

Il a été évoqué précédemment qu'il existe des modèles pour lesquels l'algorithme EM n'est pas adapté, particulièrement lorsque l'espérance de la vraisemblance complétée n'est pas calculable. C'est le cas de certains modèles avec plusieurs variables latentes, comme celui de [El Assaad et al., 2016], où le calcul de cette espérance requiert de sommer/d'intégrer sur l'ensemble des combinaisons possibles de ces variables. Afin de contourner ce problème, l'inférence variationnelle permet d'approcher le maximum de vraisemblance sans passer par le calcul direct de cette dernière. Cette méthode est détaillée dans [Blei et al., 2017] ou encore [Corduneanu and Bishop, 2001].

Considérons le cas général des données observées $\mathbf{x}$ et des variables latentes (non observées) $\mathbf{y}$ avec $\boldsymbol{\Theta}$ le vecteur de paramètres du modèle. Dans ce cas, la loi conditionnelle de l'ensemble des variables latentes $\mathbf{y}$ est définie par :

$$p(\mathbf{y}|\mathbf{x};\boldsymbol{\Theta}) = p(\mathbf{x},\mathbf{y};\boldsymbol{\Theta}) / p(\mathbf{x};\boldsymbol{\Theta}) \tag{3.25}$$

La méthode usuelle d'estimation des paramètres $\boldsymbol{\Theta}$ consiste à maximiser la log-vraisemblance

$$\mathscr{L}(\boldsymbol{\Theta}) = \log p(\mathbf{x};\boldsymbol{\Theta}) = \log \int p(\mathbf{x},\mathbf{y};\boldsymbol{\Theta})\, d\mathbf{y}. \tag{3.26}$$

Dans la situation où cette intégrale, tout comme la loi conditionnelle de $\mathbf{y}$ ne sont pas calculables, l'inférence variationnelle permet d'approcher le maximum de vraisemblance.

Borne inférieure de la log-vraisemblance et distance de Kullback-Leibler

Le principe de l'inférence variationnelle est de chercher une densité $q(\mathbf{y})$ parmi une famille de densité, notée $\mathscr{F}$ afin d'approximer la densité conditionnelle exacte. Pour cela, il faut minimiser la divergence de Kullback-Leibler de la densité candidate $q(\mathbf{y})$ par rapport à la densité $p(\mathbf{y}|\mathbf{x};\boldsymbol{\Theta})$.

Ce problème de minimisation revient à maximiser une **Borne inférieure de la log-vraisemblance**, appelée aussi ELBO et noté $F()$ qui est définie telle que :

$$\begin{aligned} F(q,\boldsymbol{\Theta}) &= \int_y q(\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y};\boldsymbol{\Theta})}{q(\mathbf{y})} d\mathbf{y} \\ &= \mathscr{L}(\boldsymbol{\Theta}) - KL(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x};\boldsymbol{\Theta})). \end{aligned} \tag{3.27}$$

Cette borne inférieure est centrale dans l'inférence variationnelle. Avec $\mathscr{F}$, une famille de densité, elle tire son nom de la relation suivante :

$$\forall q \in \mathscr{F}, F(q,\boldsymbol{\Theta}) \leq \log p(\mathbf{x};\boldsymbol{\Theta}). \tag{3.28}$$

Par conséquent, l'approximation du maximum de vraisemblance pour l'estimation des paramètres du modèle passe par la maximisation de cette borne inférieure de la log-vraisemblance.

L'algorithme d'inférence variationnelle est un algorithme itératif. Partant de valeurs initiales $q^{(0)}$ et $\boldsymbol{\Theta}^{(0)}$, l'approximation consiste à répéter successivement deux étapes jusqu'à ce que l'algorithme converge vers une solution.

Les deux étapes de l'algorithme VEM à l'itération $(q+1)$

Étape 1 Mise à jour des paramètres variationnels :

$$q^{(q+1)} = \underset{q \in \mathscr{F}}{\operatorname{argmax}}\, F(q, \boldsymbol{\Theta}^{(q)}) \tag{3.29}$$

Étape 2 Mise à jour des paramètres $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta}^{(q+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, F(q^{(q+1)}, \boldsymbol{\Theta}) \tag{3.30}$$

La difficulté réside dans la variété des formes possibles de la densité $q(.)$, appelée *densité libre*. Il est d'usage de restreindre la famille de densité $\mathscr{F}$. Lorsque la famille choisie est paramétrique, les paramètres qui définissent $q(.)$ sont appelés *paramètres variationnels*, qu'on notera $\boldsymbol{\beta}$ avec $q(\mathbf{y}) = q(\mathbf{y}; \boldsymbol{\beta})$. Par exemple, dans [Gershman et al., 2012], les auteurs proposent de ne considérer que des densités gaussiennes avec une contrainte sur les matrices de covariances. En présence d'un ensemble de variables latentes, il est d'usage d'imposer une contrainte de factorisation sur la densité $q(.)$. Par exemple, dans le cas où $\mathbf{y}$ et $\mathbf{z}$ sont des variables inobservées, alors il est souvent considéré que la famille de densité $\mathscr{F}$ est telle que $q(\mathbf{z}, \mathbf{y}) = p(\mathbf{z})\, p(\mathbf{y})$. Dans [Blei et al., 2017], les auteurs proposent de restreindre les densités à la famille exponentielle. Le choix des contraintes imposées à la famille de densité $\mathscr{F}$ dépend des données observées, des variables latentes et du modèle concerné.

Les méthodes d'inférence variationnelle permettent d'approximer le maximum de vraisemblance dans le cas où il n'est possible d'obtenir une expression littérale de la log-vraisemblance. En effet, l'algorithme EM permet d'estimer les paramètres du modèle dans le cas où il est possible d'écrire et de calculer l'espérance conditionnelle de la log-vraisemblance. Or, dans le cas de modèle trop complexe, cette étape n'est pas réalisable. C'est le cas de certains modèles de mélange à variables latentes. Parmi eux, le modèle de classification avec centres de classes dynamiques et effets régressifs présenté dans le chapitre suivant constitue un bon exemple de cas d'application de l'inférence variationnelle. Le modèle dynamique présenté dans la sous-section 3.2.3 est aussi un exemple de modèle où la vraisemblance n'est pas calculable et où l'inférence variationnelle permet d'estimer les paramètres du modèle.

Le point commun de ces méthodes d'estimation est que le nombre de composantes latentes K est un hyperparamètre. Par conséquent, il faut mettre en place des stratégies pour fixer la valeur cet hyperparamètre et plus largement sélectionner un modèle parmi un ensemble de possibilités.

3.4 La sélection de modèle

La question de la sélection de modèle est une problématique centrale dans l'approche probabiliste de classification. Comme dit précédemment, on se place dans le cadre des modèles de mélange où le nombre de composantes du mélange, K, est un hyperparamètre qui doit être défini en amont.

CADRE ET NOTATIONS

On considère un échantillon de variables indépendantes et identiquement distribuées $(\mathbf{x}_1, ...\mathbf{x}_n)$ dont les paramètres $\boldsymbol{\Theta}$ de la densité sont inconnus. On considère une collection de modèles $\mathscr{C}$ et les paramètres estimés correspondant $(\hat{\boldsymbol{\Theta}}_m)_{m\in\mathscr{C}}$. La recherche directe du modèle optimal nécessiterait de minimiser une fonction de perte entre les densités estimées $f(\hat{\boldsymbol{\Theta}}_m)$ et la densité $f(\boldsymbol{\Theta})$. Dans le cas où la densité f est inconnue, cette recherche du modèle optimal nécessite d'utiliser une autre méthode. Parmi ces méthodes, il existe des critères pénalisés basés sur la vraisemblance du modèle dont la pénalité est une fonction uniquement de la complexité du modèle. Ils seront présentés dans la première sous-section. La seconde sous-section sera dédiée à la méthode d'heuristique de pente qui est une méthode de calibration de la pénalité pour construire un critère considéré comme optimal.

3.4.1 Les critères pénalisés basés sur la vraisemblance du modèle

La sélection des modèles repose sur un compromis entre la complexité du modèle et la précision. Plus un modèle est complexe, plus il devrait être performant en matière de précision dans les estimations ou dans la classification. Il est souhaitable de limiter la complexité d'un modèle par souci de parcimonie ou afin d'éviter les problèmes de sur-apprentissage. La question de la sélection du nombre de composantes dans le cadre des modèles de mélange ou des facteurs utilisés dans le cadre de mélange de régressions, repose sur le compromis entre construire un modèle qui soit le mieux adapté aux données et construire un modèle qui contienne un nombre restreint de composantes pour éviter les problèmes de sur-apprentissage et simplifier l'analyse. Dans le cadre de l'estimation des paramètres d'un modèle par maximisation de la vraisemblance, les critères BIC, AIC et ICL ont été développés. Ces critères sont basés sur le compromis entre la vraisemblance du modèle et la complexité du modèle via une pénalité calculée en fonction du nombre de paramètres. En effet, la vraisemblance d'un modèle est croissante avec le nombre de paramètres, car plus un modèle est complexe, mieux il s'adapte aux données d'apprentissage.

PRINCIPE GÉNÉRAL D'UN CRITÈRE PÉNALISÉ

On considère une collection de modèles et les estimateurs correspondants $f(\hat{\boldsymbol{\Theta}}_m)$. Le critère basé sur la vraisemblance du modèle, dans sa forme générale, se définit tel que :

$$\mathscr{C}(m) = -\log(L_m(\mathbf{x}; \hat{\boldsymbol{\Theta}}_m)) + p(m), \tag{3.31}$$

avec L_m et $p(.)$ la vraisemblance et la pénalité du modèle m.

Le modèle m^* est considéré comme optimal selon le critère $\mathscr{C}$ s'il respecte

$$m^* = \underset{m \in \mathscr{C}}{\operatorname{argmin}} \mathscr{C}(m). \tag{3.32}$$

Il s'agit de la forme générale d'un critère pénalisé. Cependant, plusieurs auteurs ont défini des critères en spécifiant la pénalité la plus adéquate.

AKAIKE INFORMATION CRITERION (AIC)

Le critère d'information d'Akaike, introduit dans [Akaike, 1974], se base sur la vraisemblance d'un modèle m et un terme de pénalité qui dépend du nombre de paramètres libres du modèle. Il est défini tel que :

$$AIC(m) = -2\log(L_m(\mathbf{x}; \hat{\mathbf{\Theta}}_m)) + 2n_k(m), \tag{3.33}$$

avec $L_m(\mathbf{x}; \hat{\mathbf{\Theta}}_m)$ la vraisemblance du modèle après estimation du modèle m, $n_k(m)$ le nombre de paramètres. Dans [Burnham and Anderson, 2004] la comparaison des différents critères de sélection de modèle met en évidence, dans le cas d'échantillon de petite taille, que le nombre de paramètres du modèle doit rester restreint. Le critère d'information d'Akaike de second ordre (AIC_C) permet d'augmenter la contrainte liée à la complexité du modèle lorsque la taille de l'échantillon est faible. Il se définit par :

$$AIC_c(m) = AIC(m) + 2n_k(m)\left(\frac{n_k(m)+1}{n - n_k(m) - 1}\right) \tag{3.34}$$

avec $n_k(m)$ le nombre de paramètres et n la taille de l'échantillon.

BAYESIAN INFORMATION CRITERION (BIC)

Le critère du BIC, développé par [Schwarz, 1978], se base sur la vraisemblance d'un modèle m et un terme de pénalité qui dépend du nombre de paramètres libres et du nombre d'observations. Il se définit par :

$$BIC(m) = -2\log(L_m(\mathbf{x}; \hat{\mathbf{\Theta}}_m)) + n_k(m)\log(n), \tag{3.35}$$

avec $L_m(\mathbf{x}, \hat{\mathbf{\Theta}}_m)$ la vraisemblance du modèle m, $n_k(m)$ le nombre de paramètres libre du modèle et n le nombre d'observations. Le critère est décroissant en fonction de la vraisemblance et croissant en fonction de la complexité. Étant donné que l'objectif est de sélectionner un modèle qui maximise la vraisemblance en limitant la complexité, la sélection de modèle via le BIC se fait par minimisation de ce critère.

INTEGRATED CLASSIFICATION LIKELIHOOD CRITERION (ICL)

Dans le cas de modèles de classification, les critères d'information classiques ne permettent pas de sélectionner un modèle en fonction de sa capacité à classifier les observations

correctement. Dans [Biernacki et al., 1998], les auteurs suggèrent d'utiliser un autre critère spécifique aux modèles de classification. Ce critère, appelé "Integrated Classification Likelihood criterion" se définit à partir de la vraisemblance dite "complétée". En effet, la vraisemblance $L_m(\mathbf{x};\hat{\mathbf{\Theta}}_m)$ correspond à la vraisemblance associée à la distribution de $\mathbf{x}$, alors que la vraisemblance complétée est associée également à la classification latente z. Elle se définit par $L_{Cm}(\mathbf{x},\mathbf{z};\hat{\mathbf{\Theta}}_m) = \log(p(\mathbf{x},\mathbf{z};\hat{\mathbf{\Theta}}_m)$, où $\hat{\mathbf{\Theta}}_m$ désigne le vecteur de paramètres estimés du modèle m. La formule de l'ICL est la suivante :

$$ICL(m) = -2L_{C_m}(\mathbf{x},\mathbf{z};\hat{\mathbf{\Theta}}_m) + n_k(m)log(n), \tag{3.36}$$

avec $n_k(m)$ le nombre de paramètres libres du modèle et n le nombre d'observations. L'avantage de ce critère réside dans l'utilisation de la vraisemblance complète qui permet de prendre en compte la capacité du modèle à classifier les observations dans des groupes homogènes.

3.4.2 L'heuristique de pente

L'objectif de cette méthode est de définir un critère qui soit pénalisé proportionnellement à la complexité du modèle.

L'heuristique de pente est introduite dans [Birgé and Massart, 2007], où les auteurs proposent une méthode pour définir une pénalité optimale à partir d'une constante multiplicative. Cette méthode est par ailleurs détaillée dans [Baudry et al., 2012] et [Arlot, 2019]. Les auteurs passent en revue les aspects théoriques de cette méthode et proposent une mise en application de la sélection de modèle via les deux principales méthodes sous-jacente à l'heuristique de pente. On considère un échantillon de variables indépendantes et identiquement distribuées $(\mathbf{x}_1,...\mathbf{x}_n)$ de densité inconnue $f(\mathbf{x};\mathbf{\Theta})$. On considère une collection de modèles et les estimateurs correspondant $f(\mathbf{x};\hat{\mathbf{\Theta}}_m)$.

La notion de contraste dans le cas du maximum de vraisemblance

Le critère de sélection de modèle considéré dans la méthode d'heuristique de pente repose en partie sur la notion de contraste. Dans le cas du maximum de vraisemblance, la fonction de contraste $\gamma(f,\mathbf{x})$, pour un modèle m, se définit telle que :

$$\gamma(f(\mathbf{x};\hat{\mathbf{\Theta}}_m),\mathbf{x}) = -\log(f(\mathbf{x};\hat{\mathbf{\Theta}}_m)) \tag{3.37}$$

À noter que la forme du contraste dépend de la méthode d'estimation utilisée [Maugis and Michel, 2008].

Le critère pénalisé

La méthode d'heuristique de pente a pour objectif de construire un critère en calibrant la pénalité optimale. Le critère et cette pénalité se définissent par

$$\begin{cases} \text{CRIT}(m) = -\log(f(\mathbf{x}; \hat{\boldsymbol{\Theta}}_m)) + pen_{\text{OPT}}(m), \\ \\ pen_{\text{OPT}}(m) = \kappa_{\text{OPT}}\, pen_{\text{SHAPE}}(m). \end{cases} \tag{3.38}$$

Le terme $pen_{\text{SHAPE}}(m)$ est une fonction de la complexité du modèle C_m. Dans le cas de modèle de mélange, on choisit $pen_{\text{SHAPE}}(m) = C_m$. Dans ce cas, la pénalité est croissante avec la complexité du modèle, qui équivaut dans ce cas au nombre de paramètres libres du modèle, tandis que le contraste est décroissant. Dans [Baudry et al., 2012], les auteurs discutent du choix de la pénalité et de la complexité pour différents modèles.

DÉTERMINER κ_{OPT}

Toute la difficulté et l'intérêt de cette méthode résident dans la recherche d'une valeur optimale de κ_{OPT}. Pour cela, il existe deux principales méthodes. La première, basée sur les données, consiste à estimer successivement, pour différents ensembles de points, les pentes de la relation linéaire entre la log-vraisemblance et la complexité du modèle. Puis dans un second temps de construire une collection de modèles avec leur complexité associée en minimisant le critère (3.38). Ensuite, κ_{OPT} est finalement sélectionné parmi cette collection. Cette méthode est détaillée dans les travaux de [Maugis and Michel, 2011].

La seconde méthode, appelée "Saut de dimension" et présentée dans [Maugis and Michel, 2008], repose sur l'identification d'un saut dans la complexité d'une collection de modèles associés aux valeurs de κ. La justification et les preuves de l'existence d'un tel saut reposent sur des théorèmes qui ont été traités et détaillés dans [Arlot, 2019].

Pour la mise en application de cette méthode, la complexité du modèle considérée a un impact sur les résultats. Comme dit précédemment, dans le cas de modèles de mélange, on choisit comme complexité d'un modèle, le nombre de paramètres libres. Ce choix de la complexité n'est pas toujours évident et dépend des modèles considérés [Baudry et al., 2012; Arlot, 2019].

3.5 Conclusion du chapitre

Ce chapitre a présenté les méthodes et les outils qui seront réutilisés en partie par la suite. Les modèles de mélange, qui constituent une famille importante des modèles de classification non supervisée, sont appréciés pour leurs performances et pour la flexibilité qu'ils offrent. La première partie du chapitre a ainsi présenté, dans le cas de données temporelles, les modèles de mélange gaussiens, les mélanges de régressions et un modèle qui permet d'estimer des centres de classes non-stationnaires. La seconde partie du chapitre a mis en lumière les principes et la mise en œuvre d'un algorithme EM dans le cas des mélanges. Pour les cas où l'algorithme EM n'est pas adapté, les méthodes d'inférence variationnelle peuvent être utilisées pour estimer les modèles. Pour finir, la troisième partie a été dédiée aux critères et

aux méthodes de sélection de modèle. En effet, dans le cas d'une classification non supervisée, le nombre de clusters est un hyperparamètre à définir en amont de l'estimation. Nous avons fait le choix de présenter et d'utiliser par la suite les critères pénalisés du type BIC ou AIC ainsi que la méthode d'heuristique de pente.

L'ensemble de ces éléments permet d'introduire des notions utiles au chapitre 4. En effet, ce chapitre sera dédié à la présentation d'un modèle de classification visant à expliquer des données temporelles à partir de variables latentes dynamiques et d'effets régressifs liés à des variables exogènes communes. Ce modèle se place dans le cadre des modèles de mélange de régressions, car on cherche à estimer des effets régressifs, mais on considère également des centres de classes dynamiques modélisés à l'aide de processus stochastiques.

Chapitre 4

Classification et modélisation dynamique de données temporelles

Contents

4.1 Introduction

Le modèle décrit dans ce chapitre s'inscrit dans un contexte de classification de données énergétiques pour l'identification de comportements de présence, de chauffage ou de consommation d'énergie. Le manque de connaissances et la complexité de ces comportements rendent difficile la prédiction de la consommation d'énergie d'un bâtiment [De Wilde, 2014]. Parmi les méthodes les plus utilisées, on retrouve les classifications des K-means ou les méthodes hiérarchiques (CAH) par exemple. Les modèles de mélange [McLachlan and Peel, 2004] sont également très répandus, car ils offrent une plus grande flexibilité que les modèles K-means puisqu'ils contiennent moins de contraintes. Ces modèles usuels permettent de construire des clusters de comportements similaires. Dans le cas où l'on souhaite estimer des effets régressifs ou des dynamiques temporelles, il existe plusieurs variantes de modèles de mélange. Tout d'abord, on peut citer les modèles de mélange de régressions, qui classifient les données, mais estiment également les coefficients de régression spécifiques aux clusters par rapport à des facteurs connus [Desarbo and Cron, 1988; Wang et al., 2015]. De plus, dans le cas de la classification de données temporelles, il peut être intéressant d'estimer l'évolution et la dynamique des centres de classes. Dans ce cas, les modèles de mélange gaussiens à filtre de Kalman [Calabrese and Paninski, 2011; El Assaad et al., 2016] classifient, à chaque pas de temps, les données temporelles et estiment les centres des classes en fixant un *a priori* sur l'évolution de ces derniers. La modélisation des profils de classes à partir de processus autorégressifs permet de prendre en compte l'évolution et la dynamique des clusters dans le temps. C'est également l'un des objectifs des modèles proposés dans ces travaux. L'ensemble de ces modèles sont présentés en détail dans le chapitre 3.

Dans les deux modèles présentés dans ce chapitre, on considère l'ensemble des pas de temps d'une séquence pour construire les clusters. Une séquence d'observation appartient donc à un cluster. De plus, dans un premier temps, la partie régressive correspondant à l'effet exogène observé est commune à toutes les observations, contrairement aux modèles de mélange par régression qui identifient différents effets pour les différentes composantes du mélange. Ce choix de modélisation est motivé par le contexte applicatif de ces modèles. En effet, dans le cas de la classification de données mesurées au sein de logements, on considère que l'ensemble des logements d'un même bâtiment sont sujets à des effets communs comme la météo extérieure et on souhaite isoler une part non expliquée par ces effets communs qui soit relative aux comportements de présence ou de chauffage. Dans un second temps, le modèle dynamique proposé permettra d'estimer les profils de classes dynamiques et des effets exogènes propres à chaque cluster. L'approche adoptée ici est celle des modèles à variables latentes utilisant des modèles de mélange. Afin d'estimer les paramètres des modèles présentés, des méthodes d'inférence variationnelle sont utilisées pour approcher l'optimum de la vraisemblance.

L'idée sous-jacente du modèle proposé est de considérer que des variables mesurées

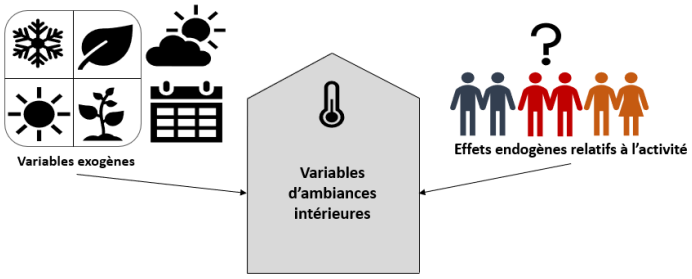

FIGURE 4.1 – Schéma d'introduction au modèle proposé. On considère que les variables mesurées au sein des logements d'un immeuble sont le résultat d'effet de facteurs exogènes et connus tels que les variables météorologiques et calendaires, et d'effets endogènes relatifs à l'activité et à la présence.

au sein d'un appartement sont le résultat d'effets exogènes, relatifs à des facteurs connus et mesurables comme les conditions extérieures, et d'une partie restante dite endogène (4.1). La partie exogène, fait référence à des effets communs à l'ensemble des logements qui dépendent de variables contextuelles et calendaires. La partie endogène, relative aux comportements de chauffage des habitants, à leur présence et leurs activités, est un sujet important dans le domaine de l'énergie. Dans un contexte de classification, on considère que ces comportements peuvent être estimés sous forme de clusters afin d'identifier et de résumer les comportements types.

La première section de ce chapitre est dédiée à la présentation du modèle de classification à profils dynamiques et effet régressif commun. La seconde section traite de l'estimation des paramètres et de l'algorithme utilisé. La troisième section est dédiée à l'évaluation des performances du modèle à partir de jeux de données simulées statistiquement. Pour finir, la dernière section propose une extension du modèle avec une forme moins contrainte et des effets exogènes estimés pour chaque cluster. Ce second modèle sera évalué à partir de différents jeux de données.

4.2 Construction d'un modèle de classification à profils dynamiques et effet régressif commun

L'objectif est de construire un modèle de classification pour des séries temporelles sujettes à des effets exogènes connus et communs à l'ensemble des observations. Ces effets sont considérés comme communs pour les différencier des effets endogènes propres à chaque cluster. Les centres de classes sont considérés comme des processus dynamiques et seront modélisés comme tels.

Formalisme

Pour formaliser ce modèle, n représente le nombre d'entités et T représente le nombre de pas de temps observés. Ensuite, nous considérons les éléments suivants :

- $(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n)$ un ensemble de n observations avec $\forall i \in [\![1; n]\!],\ \mathbf{x}_i = (x_{i1}, \ldots, x_{it}, \ldots, x_{iT})$, où $\forall t \in [\![1, T]\!],\ x_{it} \in \mathbb{R}$,
- $\forall t \in [\![1; T]\!], \mathbf{u}_t \in \mathbb{R}^p$ un vecteur représentant p facteurs exogènes et observables tels que des données météorologiques ou des variables calendaires. Ces facteurs sont ici supposés communs à l'ensemble des entités indicées par i.

Dans le cas où la température intérieure serait observée, il est possible d'expliquer cette température à l'aide d'un ensemble de facteurs exogènes et d'une partie non observée représentant le comportement des occupants. Supposons qu'il existe K classes non observées, on considère que l'observation $x_{it} \in \mathbb{R}$ peut être expliquée par le modèle suivant :

$$\forall i \in [\![1; n]\!],\ \ \forall t \in [\![1; T]\!];\quad x_{it} = \mathbf{u}_t' \mathbf{a} + \sum_{k=1}^{K} z_{ik} b_{kt} + e_{it}, \tag{4.1}$$

où z_{ik} est une variable binaire égale à 1 si l'observation i appartient à la classe k et 0 sinon. On peut alors introduire la variable z_i qui vaut k si z_{ik} vaut 1, avec $z_i \sim \mathcal{M}(\boldsymbol{\pi})$ et $\boldsymbol{\pi} = (\pi_k)_{k=1,\ldots,K}$ désignant les proportions du mélange. De plus, le processus $(b_{kt})_{t=1,\ldots,T}$ désigne le profil de la classe k, e_{it} est un terme de bruit supposé gaussien, centré et de variance $v_{z_i}^2$ et $\mathbf{a} \in \mathbb{R}^p$ désigne le vecteur de coefficients associés aux facteurs exogènes.

Un des objectifs de ce modèle est de prendre en compte et de modéliser la dynamique des profils de classes. Des processus dynamiques permettent de caractériser les classes comme suit :

$$\forall t \in [\![1, T]\!]; \forall k \in [\![1, K]\!];\quad b_{kt} = \Phi_k b_{kt-1} + \nu_{kt}, \tag{4.2}$$

avec $b_{k0} \sim \mathcal{N}_d(\mu_{k0}, \sigma_{k0}^2)$ et ν_{kt} un bruit gaussien tel que $\nu_{kt} \sim \mathcal{N}_d(0, w_k^2)$. Grâce aux éléments précédents, le vecteur des paramètres du modèle sera noté $\boldsymbol{\Theta}$ tel que :

$$\boldsymbol{\Theta} = \left\{(v_k^2, w_k^2, \pi_k, \Phi_k, \mu_{k0}, \sigma_{k0}^2)_{k=1,\ldots K}, \mathbf{a}\right\}, \tag{4.3}$$

Les profils de classes sont modélisés comme des processus autorégressifs de premier ordre. Ce choix de modélisation des processus latents pour caractériser les classes a déjà été proposé dans [Calabrese and Paninski, 2011] et [El Assaad et al., 2016]. Ces auteurs ont également choisi cet *a priori* sur les centres de classes dans le cadre de modèles de classification dynamique. Ils montrent que ces modèles ont de meilleures performances sur des jeux de données simulées que d'autres modèles tels que les mélanges de régressions ou les modèles de mélange gaussien plus simples. De plus, il faut noter que la modélisation basée sur des processus autorégressifs d'ordre un est parcimonieuse, limitant le nombre de paramètres à estimer. De plus, ce choix offre la possibilité de prédire, à court terme, le comportement futur des profils $(b_{kt})_t$. D'autres

modèles de processus latents peuvent cependant être envisagés. Par exemple, dans [Samé et al., 2009], les auteurs caractérisent les classes de polynômes dépendant du temps dans un contexte de régression non linéaire avec différents régimes.

Identifiabilité du modèle et contraintes

Le modèle défini par l'équation (4.1) n'est pas identifiable. En effet, le coefficient a_0 peut être confondu avec les profils de classes $(b_{kt})_{(k,t)}$. Dans ce cas, il est nécessaire d'ajouter une contrainte au modèle. Dans le cas présent, en fixant $\tilde{\mathbf{a}} = (a_1, \ldots, a_p)$, et en notant $\tilde{\mathbf{u}}_t$ les variables exogènes à p dimensions correspondantes, on a :

$$\mathbf{u}_t'\mathbf{a} + \sum_k z_{ik} b_{kt} = a_0 + \tilde{\mathbf{u}}_t'\tilde{\mathbf{a}} + \sum_k z_{ik} b_{kt} = (a_0 - \alpha) + \tilde{\mathbf{u}}_t'\tilde{\mathbf{a}} + \sum_k z_{ik}(b_{kt} + \alpha). \tag{4.4}$$

Ainsi, en fonction de la valeur de α, il existe un nombre infini de choix pour a_0 et b_{kt}. Pour assurer l'identifiabilité, nous ajoutons la contrainte suivante au modèle : $\sum_{k=1}^{K}\sum_{t=1}^{T} b_{kt} = 0$.

La section suivante est consacrée aux méthodes et aux algorithmes d'estimation, en particulier aux méthodes d'inférence variationnelle.

4.3 Estimation des paramètres du modèle

4.3.1 Inférence variationnelle et Borne inférieure de la log-vraisemblance

L'estimation des paramètres du modèle pourrait être un problème classique de maximisation de la log-vraisemblance. La log-vraisemblance peut être écrite comme suit :

$$\begin{aligned} \mathscr{L}(\mathbf{x};\boldsymbol{\Theta}) &= \log p(\mathbf{x};\boldsymbol{\Theta}) \\ &= \log\left(\sum_z p(\mathbf{z};\boldsymbol{\Theta}) \int_{\mathbf{b}} p(\mathbf{b} \mid \mathbf{z};\boldsymbol{\Theta})\, p(\mathbf{x} \mid \mathbf{b},\mathbf{z};\boldsymbol{\Theta})\, d\mathbf{b}\right) \end{aligned} \tag{4.5}$$

Mais dans ce cas, la dépendance temporelle des variables latentes $(\mathbf{b},\mathbf{z})$ rend la fonction précédente non calculable. Il est alors nécessaire de contourner ce problème en utilisant des méthodes d'inférence variationnelle. Cette méthode, détaillée dans le chapitre 3, permet d'approcher le maximum de vraisemblance en maximisant une borne inférieure à l'aide d'un algorithme itératif.

Cette "Borne inférieure de la log-vraisemblance" (ELBO), notée $F(q,\boldsymbol{\Theta})$ est construite à partir d'une distribution variationnelle, noté $q(.)$. Pour rappel, l'inférence variationnelle est basée sur la recherche d'une distribution $q(.)$, issue d'une famille $\mathscr{F}$, qui minimise la divergence de Kullback-Leibler de cette densité par rapport à la loi de probabilité *a posteriori*

des variables latentes du modèle. Dans le chapitre 3 on a défini $F(.)$ telle que :

$$\begin{aligned} F(q,\boldsymbol{\Theta}) &= \int_{\mathbf{z},\mathbf{b}} q(\mathbf{z},\mathbf{b}) \log \frac{p(\mathbf{x},\mathbf{z},\mathbf{b};\boldsymbol{\Theta})}{q(\mathbf{z},\mathbf{b})} d\mathbf{z}d\mathbf{b} \\ &= \int_{\mathbf{z},\mathbf{b}} q(\mathbf{z},\mathbf{b}) \log p(\mathbf{z},\mathbf{b},\mathbf{x};\boldsymbol{\Theta}) d\mathbf{z}d\mathbf{b} - \int_{\mathbf{z},\mathbf{b}} q(\mathbf{z},\mathbf{b}) \log q(\mathbf{z},\mathbf{b}) d\mathbf{z}d\mathbf{b} \\ &= \mathbf{E}_q(\mathscr{L}_c(\boldsymbol{\Theta})) + H(q), \end{aligned} \tag{4.6}$$

où $H(q)$ est l'entropie de la fonction de densité $q(.)$, et $\mathscr{L}_c$ désigne la log-vraisemblance complétée du modèle définies par :

$$H(q) = -\int_{\mathbf{z},\mathbf{b}} q(\mathbf{z},\mathbf{b}) \log q(\mathbf{z},\mathbf{b}) d\mathbf{z}d\mathbf{b} \tag{4.7}$$

$$\begin{aligned} \mathscr{L}_c(\boldsymbol{\Theta}) &= \mathscr{L}(\mathbf{x},\mathbf{z},\mathbf{b};\boldsymbol{\Theta}) = \log p(\mathbf{x},\mathbf{z},\mathbf{b};\boldsymbol{\Theta}) \\ &= \log(p(\mathbf{x},\mathbf{z} \mid \mathbf{b};\boldsymbol{\Theta})) + \log(p(\mathbf{b};\boldsymbol{\Theta})) \\ &= \sum_{i,k,t} \log(p(x_{it}, z_{ik} \mid \mathbf{b};\boldsymbol{\Theta})) + \sum_{k,t} \log(p(b_{kt} \mid b_{kt-1};\boldsymbol{\Theta})) + \sum_{k} \log(p(b_{k0};\boldsymbol{\Theta})) \\ &= \sum_{i,k,t} \log(\pi_k \varphi(x_{it}, b_{kt} + \mathbf{u}_t'\mathbf{a}, v_k^2)) + \sum_{k,t} \log(\varphi(b_{kt}, \Phi_{kt-1}, w_k^2)) + \sum_{k} \log(\varphi(b_{k0}, \mu_{k0}, \sigma_{k0}^2)). \end{aligned} \tag{4.8}$$

L'objectif principal est d'estimer la fonction de densité variationnelle et les paramètres du modèle en maximisant la borne inférieure. Afin de simplifier la recherche de la solution pour la fonction $q(.)$ et assurer une solution, il est possible de réduire la forme possible de la densité variationnelle à une famille de fonctions restreintes. Dans [Blei et al., 2017], les auteurs justifient le choix de la "mean-field family", car cela permet de simplifier le problème d'optimisation tout en offrant de bonnes performances. La famille du champ moyen correspond à l'hypothèse de factorisation de la distribution q :

$$q(\mathbf{z},\mathbf{b}) = \prod_{i=1}^{n} q_z(z_i) \prod_{t=0}^{T} \prod_{k=1}^{K} q_b(b_{kt}), \tag{4.9}$$

où q_z est la distribution de la variable latente z_i et q_b est la distribution des processus (b_{kt}). Dans ce modèle, les variables b_{kt} sont supposées gaussiennes avec des paramètres de moyenne m_{kt} et de variance λ_k. Les variables z_i sont distribuées selon une fonction de densité multinomiale de paramètres $(\tau_{ik})_{i=1,\dots n;k=1,\dots,K}$. La fonction $q(.)$ peut être réécrite telle que :

$$q(\mathbf{z},\mathbf{b}) = q(\mathbf{m},\boldsymbol{\tau},\boldsymbol{\lambda}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \tau_{ik}^{z_{ik}} \prod_{t=0}^{T} \prod_{k=1}^{K} \varphi(b_{kt}, m_{kt}, \lambda_k). \tag{4.10}$$

Cette fonction de densité variationnelle nous amène à introduire les paramètres variationnels qui seront estimés en maximisant la borne inférieure de la log-vraisemblance. Les paramètres variationnels des modèles sont les suivants :

- $\boldsymbol{\tau} = \{(\tau_{ik})_{k=1,..,K;i=1,..,n}\}$
- $\mathbf{m} = \{(m_{kt})_{k=1,..,K;t=0,..,T}\}$
- $\boldsymbol{\lambda} = \{(\lambda_k)_{k=1,..,K}\}$.

Selon le modèle 4.1, les formules 4.6 et 4.10, il est possible d'écrire explicitement la borne inférieure de la log-vraisemblance telle que :

$$\begin{aligned} F(\mathbf{m},\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\Theta}) = & \sum_{i,t,k} \tau_{ik}\left(\log(\pi_k\varphi(x_{it};m_{kt}+\mathbf{u}_t'\mathbf{a},v_k^2)) - \frac{1}{2}\lambda_k(v_k^{-2})\right) \\ & + \sum_{k,t}\log\left(\varphi(m_{kt};\Phi_k m_{kt-1},w_k^2)\right) - \frac{1}{2}\lambda_k((w_k^{-2}) + (w_k^{-2}\Phi_k^2)) \\ & + \sum_{k}\log\left(\varphi(m_{k0},\mu_{k0},\sigma_{k0}^2)\right) - \frac{1}{2}\lambda_k(\sigma_{k0}^{-2}) \\ & - \sum_{i,k}\tau_{ik}\log(\tau_{ik}) + \frac{d(T+1)}{2}\sum_k \log(2\pi e) + \log(\lambda_k). \end{aligned} \tag{4.11}$$

Par conséquent, l'estimation des paramètres du modèle peut être obtenue en itérant les deux étapes suivantes :

- Approximer les paramètres variationnels : $(\boldsymbol{\tau}^*,\mathbf{m}^*,\boldsymbol{\lambda}^*) = \underset{\boldsymbol{\tau},\mathbf{m},\boldsymbol{\lambda}}{\operatorname{argmax}}(F(\boldsymbol{\tau},\mathbf{m},\boldsymbol{\lambda},\boldsymbol{\Theta}))$.
- Estimer les paramètres du modèle : $\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}(F(\boldsymbol{\tau}^*,\mathbf{m}^*,\boldsymbol{\lambda}^*,\boldsymbol{\Theta}))$.

L'algorithme utilisé est itératif et consiste en la mise à jour de chaque paramètre, un par un, en considérant les autres comme fixés.

4.3.2 Algorithme Variational-Expectation-Maximisation (VEM)

La section précédente nous amène à construire un algorithme qui maximise itérativement la borne inférieure en fonction de chaque paramètre, un par un, en considérant les autres comme fixes. Pour rappel, les paramètres des modèles sont désignés par $\boldsymbol{\Theta}$ tels que : $\boldsymbol{\Theta} = \left\{(v_k^2,w_k^2,\pi_k,\Phi_k,\mu_{k0},\sigma_{k0}^2)_{k=1,...K},\mathbf{a}\right\}$, et les paramètres variationnels sont $(\boldsymbol{\tau},\mathbf{m},\boldsymbol{\lambda}) = \left\{(\tau_{ik},m_{kt},\lambda_k)_{k=1,...,K}\right\}$, L'algorithme est construit en deux parties principales. La première consiste en l'initialisation de tous les paramètres qui seront estimés. La seconde partie, appelée "itération" correspond à la partie itérative où, de façon alternée, les paramètres variationnels et les paramètres du modèle sont mis à jour jusqu'à ce que le critère d'arrêt soit atteint.

Initialisation

Les premiers paramètres à initialiser sont les variances $(w_k^2,v_k^2,\sigma_{k0}^2)$, les variances variationnelles (λ_k) et les proportions de classe (π_k). Dans l'algorithme proposé, les variances initiales $(w_k^{2(0)},v_k^{2(0)},\lambda_k^{(0)})$ sont fixées à 1, ce qui est un choix classique. Concernant les paramètres $(\sigma_{k0}^{2(0)})$, une valeur élevée est choisie pour l'initialisation en raison du manque

d'informations. Pour les résultats présentés dans ce chapitre, les variances ($\sigma_{k0}^{2(0)}$) sont fixées à 2. Ensuite, les paramètres de proportion ($\pi_k^{(0)}$) sont fixés à $1/K$.

L'initialisation des paramètres ($\mathbf{m}, \boldsymbol{\tau}, \mathbf{a}$) est effectuée de manière itérative. Tout d'abord, les centres de classe ($m_{kt}^{(0)}$) sont fixés à 0, et les données ($\mathbf{x}_1, ..., \mathbf{x}_n$) sont partitionnées en utilisant l'algorithme K-means. Cette classification définit les probabilités initiales d'appartenance aux clusters ($\tau_{ik}^{(0)}$). En considérant les facteurs exogènes ($\mathbf{u}_t$), les centres de classe initiaux ($m_{kt}^{(0)}$), et les probabilités *a posteriori* initiales ($\tau_{ik}^{(0)}$), les coefficients de régression sont estimés à l'aide de la formule suivante, qui correspond à la résolution d'un problème de régression linéaire, pondérée par les probabilités τ_{ik} :

$$\mathbf{a}^{(0)} = \left[n \sum_t \left(\mathbf{u}_t' \mathbf{u}_t \right) \right]^{-1} \left[\sum_{t,i,k} \tau_{ik}^{(0)} \mathbf{u}_t \left(x_{it} - m_{kt}^{(0)} \right) \right]. \tag{4.12}$$

D'une seconde classification K-means effectuée sur les observations retraitées $x_{it}^* = x_{it} - \mathbf{u}_t' \mathbf{a}^{(0)}$, résulte une initialisation plus fine des centres de clusters et des probabilités *a posteriori*. Ce processus d'initialisation, pour ($\mathbf{m}^{(0)}, \boldsymbol{\tau}^{(0)}, \mathbf{a}^{(0)}$), est répété deux fois.

Pour finir, les coefficients (Φ_k) et les paramètres (m_{k0}, μ_{k0}) sont initialisés à l'aide des formules suivantes, obtenues en maximisant la borne inférieure et considérant les autres paramètres fixés aux valeurs ci-dessus :

$$\Phi_k^{(0)} = \left(\sum_{t=1}^{T} m_{kt}^{(0)} m_{kt-1}^{(0)} \right) \left(\sum_{t=1}^{T} (m_{kt}^{(0)^2} + \lambda_k^{(0)}) \right)^{-1}, \tag{4.13}$$

$$m_{k0}^{(0)} = \frac{1}{t_I} \sum_{t=1}^{t_I} m_{kt}^{(0)}, \tag{4.14}$$

$$\mu_{k0}^{(0)} = m_{k0}^{(0)}, \tag{4.15}$$

avec $t_I \leq T$. Dans notre cas, nous avons choisi $t_I = min(T, 10)$.

Il est important de noter que, si les données s'étendent sur moins de 10 instants, les derniers paramètres sont initialisés en utilisant la séquence entière. Enfin, l'algorithme est lancé avec les paramètres initiaux $\boldsymbol{\Theta}^{(0)}$, et les paramètres variationnels initiaux ($\mathbf{m}^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\tau}^{(0)}$).

Itérations

Notons (c) l'indice de l'itération courante. La mise à jour de tous les paramètres se fait selon les étapes décrites ci-dessous. Toutes les formules suivantes résultent d'un problème de maximisation de la borne inférieure présentée dans l'équation (**??**.

1. **MISE À JOUR DES PARAMÈTRES VARIATIONNELS DES CENTRES DE CLASSE** $\left(m_{kt}^{(c+1)} \right)_{(k,t)}$: La résolution du problème de maximisation n'est pas simple en raison de la dépendance temporelle de chaque centre de classe. En effet, pour estimer les paramètres $\mathbf{m}$, il

faudrait maximiser la fonction suivante :

$$W(\mathbf{m}) = \sum_{i,t,k} \tau_{ik} \left(\log(\pi_k \varphi(x_{it}; m_{kt} + \mathbf{u}_t' \mathbf{a}, v_k^2))\right) + \sum_{k,t} \log\left(\varphi(m_{kt}; \Phi_k m_{kt-1}, w_k^2)\right). \quad (4.16)$$

Cependant, la maximisation de cette fonction n'est pas directement réalisable. Pour cette raison, une version adaptée des équations du filtre de Kalman est utilisée [El Assaad et al., 2016]. Une manière alternative de maximisation de la fonction $F()$, par rapport à $(m_{kt})_{t=1,\dots,T}$, consiste à considérer le système suivant :

$$\begin{cases} m_{kt} = \Phi_k m_{kt-1} + \eta_{kt} \quad \eta_{kt} \sim \mathcal{N}(0, w_k^2) \\ x_{it} = m_{kt} + \mathbf{u}_t' \mathbf{a} + e_{it} \quad \text{Avec} \;\; e_{it} \sim \mathcal{N}(0, \frac{1}{\tau_{ik}} v_k^2) \end{cases} \quad (4.17)$$

Les équations du filtre de Kalman peuvent être divisées en deux étapes : les équations Forward et les équations Bakward. La première étape consiste à estimer les quantités d'intérêt un pas de temps après l'autre, à partir des quantités précédemment estimées. La deuxième étape, le lissage arrière, consiste à lisser les estimations précédentes en tenant compte de toutes les observations.

Équations Forward : l'initialisation de l'algorithme du filtre de Kalman est donnée par :

$$\begin{cases} c_{k0} = \mu_{k0}^{(c)} \\ \mathbf{P}_{k0} = \sigma_{k0}^{2(c)} \end{cases} \quad (4.18)$$

Puis, à partir des valeurs initiales précédentes, pour chaque $t \in \{1, \dots T\}$, les quantités suivantes sont calculées :

$$\forall k = 1, \dots, K; t = 1, \dots T \quad \begin{cases} \mathbf{P}_{kt} = \left(\left(\Phi_k^{(c)} \mathbf{P}_{kt-1} \Phi_k^{'(c)} + w_k^{2(c)}\right)^{-1} + \sum_i (\tau_{ik})^{-1(c)} v_k^{2(c)}\right)^{-1} \\ c_{kt} = \Phi_k^{(c)} c_{kt-1} + \mathbf{P}_{kt} v_k^{-2(c)} \sum_i \tau_{ik}^{(c)} \left(x_{it} - \mathbf{u}_t' \mathbf{a}^{(c)} - \Phi_k^{(c)} c_{kt-1}\right) \end{cases} \quad (4.19)$$

Équation Backward : la deuxième étape consiste à réajuster les valeurs estimées en utilisant toutes les données disponibles. Ensuite, pour $t \in \{T-1, \dots, 0\}$ on a :

$$\forall k = 1, \dots, K; \; t = T-1, \dots, 0 \quad \begin{cases} m_{kt}^{(c+1)} = c_{kt} + \mathbf{L}_{kt}(m_{kt-1}^{(c+1)} - c_{kt-1}) \\ \mathbf{L}_{kt} = \mathbf{P}_{kt} \Phi_k^{'(c)} \left(\Phi_k^{(c)} \mathbf{P}_{kt} \Phi_k' + w_k^{2(c)}\right)^{-1} \end{cases} \quad (4.20)$$

et pour le premier terme avec $t = T$, les éléments lissés sont tels que : $m_{kt}^{(c+1)} = c_{kt}$. Après ce calcul, les centres de classe estimés sont centrés comme cela a été expliqué dans la section 4.2.

2. **MISE À JOUR DES PROBABILITÉS *a posteriori* D'APPARTENANCE À LA CLASSE** $\left(\tau_{ik}^{(c+1)}\right)_{(i,k)}$
En maximisant la fonction (4.11) selon τ_{ik}, $\forall i \in \{1, \dots, n\}$ et $k \in \{1, \dots, K\}$, la valeur mise

à jour est telle que :

$$\tau_{ik}^{(c+1)} = \frac{\exp\left(\sum_t \log\left(\pi_k^{(c)} \varphi(x_{it}, m_{kt}^{(c)} + \mathbf{u}_t' \mathbf{a}^{(c)}, v_k^{2(c)})\right) - \frac{1}{2} v_k^{-2(c)} \lambda_k^{(c)}\right)}{\sum_{l=1}^{K} \exp\left(\sum_t \log\left(\pi_l^{(c)} \varphi(x_{it}, m_{lt}^{(c)} + \mathbf{u}_t' \mathbf{a}^{(c)}, v_l^{2(c)})\right) - \frac{1}{2} v_l^{-2(c)} \lambda_l^{(c)}\right)} \tag{4.21}$$

3. **MISE À JOUR DES PARAMÈTRES DE VARIANCES VARIATIONNELLES** $\left(\lambda_k^{(c+1)}\right)_{(k)}$ En maximisant l'ELBO selon les paramètres des variances variationnelles, nous obtenons la formule suivante :

$$\forall k = 1, \ldots, K, \quad \lambda_k^{(c+1)} = \frac{(T+1)d}{T v_k^{-2(c)} \sum_i \tau_{ik}^{(c+1)} + T w_k^{-2(c)} (1 + \Phi_k^{2(c)})) + \sigma_{k0}^{-2(c)}} \tag{4.22}$$

4. **MISE À JOUR DES PROPORTIONS DE CLASSE** $\left(\pi_k^{(c+1)}\right)_{(k)}$ Afin d'obtenir un modèle plus parcimonieux et d'éviter l'estimation de clusters vides, ces paramètres sont fixés à une valeur habituelle :

$$\forall k \in \{1, \ldots, K\}, \pi_k = \frac{1}{K} \tag{4.23}$$

Il s'agit d'une hypothèse forte, mais qui est également utilisée par d'autres méthodes de classification telles que l'algorithme K-means. La modélisation sous-jacente à la classification K-means est un sous-cas du mélange gaussien avec certaines hypothèses dont l'égalité des proportions des classes comme cela fût précisé précédemment. En outre, il est important de noter que ces paramètres peuvent être considérés comme des proportions de classes *a priori* car, à la fin, la classification *a posteriori* est obtenue sur la base des probabilités (τ_{ik}).

5. **MISE À JOUR DES PARAMÈTRES** $\left(\mu_{k0}^{(c+1)}, \Sigma_{k0}^{(c+1)}\right)_{(k)}$ La mise à jour de ces paramètres est telle que :

$$\begin{cases} \mu_{k0}^{(c+1)} = m_{k0}^{(c+1)} \\ \\ \sigma_{k0}^{2(c+1)} = \lambda_k^{(c+1)} \end{cases} \tag{4.24}$$

6. **MISE À JOUR DES PARAMÈTRES DES PROCESSUS** $\left(\Phi_k^{(c+1)}\right)_{(k)}$ La mise à jour des paramètres des processus est telle que :

$$\Phi_k^{(c+1)} = \left(\sum_t m_{kt}^{(c+1)} m_{kt-1}^{'(c+1)}\right) \left(\sum_t (m_{kt-1}^{(c+1)} m_{kt-1}^{'(c+1)}) + \lambda_k^{(c+1)} T\right)^{-1} \tag{4.25}$$

Une contrainte supplémentaire est ajoutée pour assurer la stationnarité des processus représentés par $(b_{kt})_{t=1,\ldots,T}$:

$$\forall k = 1, \ldots, K \quad |\phi_k^{(c+1)}| < 1 \tag{4.26}$$

7. **MISE À JOUR DES VARIANCES DES PROFILS DE CLASSES** $\left(w_k^{2(c+1)}\right)_{(k)}$ La formule de mise à jour est donnée par :

$$w_k^{2(c+1)} = \frac{1}{T}\sum_t \left(m_{kt}^{(c+1)} - \Phi_k^{(c+1)} m_{kt-1}^{(c+1)}\right)^2 + \lambda_k\left(1+\Phi_k^{2(c+1)}\right) \tag{4.27}$$

8. **MISE À JOUR DES VARIANCES** $\left(v_k^{2(c+1)}\right)_{(k)}$

   La formule de mise à jour est donnée par :

$$v_k^{2(c+1)} = \frac{1}{\sum_{i,t}\tau_{ik}^{(c+1)}}\sum_{i,t}\tau_{ik}^{(c+1)}\left(\left(x_{it} - m_{kt}^{(c+1)} - \mathbf{u}_t'\mathbf{a}^{(c)}\right)^2 + \lambda_k^{(c+1)}\right) \tag{4.28}$$

9. **MISE À JOUR DES COEFFICIENTS ASSOCIÉS AUX FACTEURS EXOGÈNES** $\mathbf{a}$

   Enfin, afin de mettre à jour les paramètres des facteurs exogènes, la formule suivante est utilisée :

$$\mathbf{a}^{(c+1)} = \left[\sum_{t,i,k}\tau_{ik}\left(\frac{\mathbf{u}_t\mathbf{u}_t'}{v_k^{2(c+1)}}\right)\right]^{-1}\left[\sum_{t,i,k}\tau_{ik}^{(c+1)}\left(\frac{1}{v_k^{2(c+1)}}\mathbf{u}_t\left(x_{it} - m_{kt}^{(c+1)}\right)\right)\right] \tag{4.29}$$

CRITÈRE D'ARRÊT

Cet algorithme itératif procède en mettant à jour à chaque itération les paramètres jusqu'à obtenir une convergence vers une solution finale. L'objectif de cet algorithme est de se rapprocher du maximum de la fonction F (4.6). Ici, on suppose que l'algorithme a convergé vers une solution lorsque les centres de classe mis à jour sont quasi identiques à ceux obtenus lors de l'itération précédente. En d'autres termes, le critère d'arrêt de cet algorithme est :

$$\frac{1}{KT}\sum_{t,k}\left(m_{kt}^{(c+1)} - m_{kt}^{(c)}\right)^2 < \epsilon, \text{ avec } \epsilon \to 0 \tag{4.30}$$

Une fois ce critère atteint, l'algorithme s'arrête.

LA CLASSIFICATION DES DONNÉES *a posteriori*

L'objectif d'un modèle de classification est de pouvoir attribuer à chaque observation une classe d'appartenance. L'algorithme présenté ici permet de calculer des probabilités *a posteriori* d'appartenance aux classes pour chaque observation (τ_{ik}). Afin d'obtenir une partition des observations sous forme de cluster, il faut ajouter une étape supplémentaire appelée **règle du Maximum *a posteriori*** . Il s'agit d'une méthode standard d'attribution des classes qui consiste à attribuer la classe pour laquelle la probabilité *a posteriori* est la plus grande. En notant $\hat{z}_i$ la classe d'appartenance estimée et $(\hat{\tau}_{ik})_k$ le vecteur de probabilités *a posteriori* estimé pour l'observation i, on a :

$$\hat{z}_i = \underset{k}{\text{argmax}}\,(\hat{\tau}_{ik})_{k=1,..K}. \tag{4.31}$$

L'ensemble de ces formules et plus généralement la méthode d'inférence variationnelle ont permis de construire l'algorithme qui sera utilisé pour l'estimation du modèle proposé. Cet algorithme est détaillé ci-dessous.

L'ALGORITHME VEM POUR LE MODÈLE PROPOSÉ

Entrée : Observations $(\mathbf{x}_1, ..., \mathbf{x}_n)$, avec $\mathbf{x}_i \in \mathbb{R}^T$, le nombre de clusters K et les vecteurs de facteurs exogènes $(\mathbf{u}_t)_{t=1,...T}$

Sortie : $(\mathbf{m}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Theta})$

Initialisation:

$\boldsymbol{\Theta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\mathbf{m}^{(0)}$, $\boldsymbol{\tau}^{(0)}$ à partir de l'algorithme K-means et de valeurs initiales définies ;

repeat

  for $k = 1$ **to** K **do**

    for $t = 1$ **to** T **do**

      Calcul des profils de classes $\left(m_{kt}^{(c)}\right)$;

      for $i = 1$ **to** n **do**

        Calcul des probabilités d'appartenance $\left(\tau_{ik}^{(c)}\right)$

      end

    end

    Calcul des variances variationelles (λ_k);

    Calcul des paramètres $\left(\mu_{k0}^{(c)}, \sigma_{k0}^{(c)}\right)$;

    Calcul des proportions de classes $\left(\pi_k^{(c)}\right)$;

    Calcul des variances $\left(v_k^{(c)}\right)$ et $\left(w_k^{(c)}\right)$;

  end

  Calcul des coefficients de régression $\mathbf{a}$;

until *La critère d'arrêt est atteint*;

Algorithm 1: Algorithme d'inférence variationelle pour l'estimation du modèle

4.4 Évaluation des performances sur des données simulées

Afin d'évaluer les performances du modèle, il est important de simuler différents jeux de données. De plus, nous définissons trois critères pour évaluer la précision du modèle proposé. Ces résultats sont comparés aux performances de deux autres modèles utilisés comme références.

4.4.1 La simulation des données à partir du modèle proposé

La simulation d'un ensemble de données peut être décomposée en cinq étapes. Un exemple de données générées est présenté dans la figure 4.3.

- On fixe les paramètres du modèle $\boldsymbol{\Theta}$ avec les valeurs détaillées dans le tableau 4.1

- Pour un nombre donné de clusters K et une longueur de séquence T, les profils de classes sont générés sous forme de processus autorégressifs de premier ordre : ces profils sont simulés plus ou moins distinctement selon le niveau de difficulté choisi, puis centrés (voir Figure 4.3(B)).
- Pour chaque observation, indicée par i, les proportions d'appartenance (τ_{ik}) sont générées avec une classe majoritaire selon les paramètres π_k fixés. L'observation est le résultat d'une somme pondérée par ces proportions d'appartenances des profils de classes générés plus un bruit gaussien de paramètre $v^2_{z_i}$. Notons que z_i désigne la classe d'appartenance de l'observation i correspondant à la classe majoritaire. Il est possible de simuler des cas plus ou moins difficiles en jouant sur les proportions d'appartenances afin de générer des classes plus ou moins mixées (voir la Figure 4.9).
- À partir de facteurs exogènes connus et observables, centrés et réduits (4.2), il faut définir les coefficients associés afin de calculer l'effet régressif commun (4.3(A)). Dans le cas des simulations présentées par la suite, le choix a été fait d'utiliser des données réelles de température et d'humidité centrées réduites. Il aurait été tout à fait possible de générer ces facteurs sans se baser sur des données réelles.
- En utilisant la formule (4.1) et les éléments simulés précédents, chaque observation est générée (voir Figure 4.3(C)).

TABLE 4.1 – Paramètres utilisés pour la simulation du jeu de données.

	(π_k)	(μ_{k0})	(σ^2_{k0})	(v^2_k)	(w^2_k)	(Φ_k)	**a**
K=2	(0.5, 0.5)	(4.5, 6)	(1, 1)	(2, 2)	(1, 1)	(0.99, 0.95)	(2.5, 1.8)

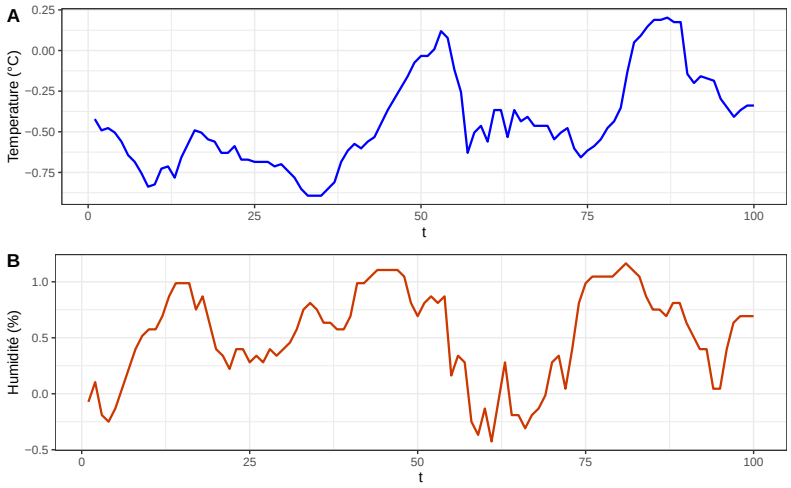

FIGURE 4.2 – Données météorologiques utilisées comme facteurs exogènes lors de la simulation des données. À noter que ces données sont centrées et réduites pour la simulation.

Les données d'entrée, qu'on appelle "observations", sont les séquences $\mathbf{x}_i$ générées et tracées sur le graphique 4.3(C). L'application de l'algorithme à ces données permet d'identifier

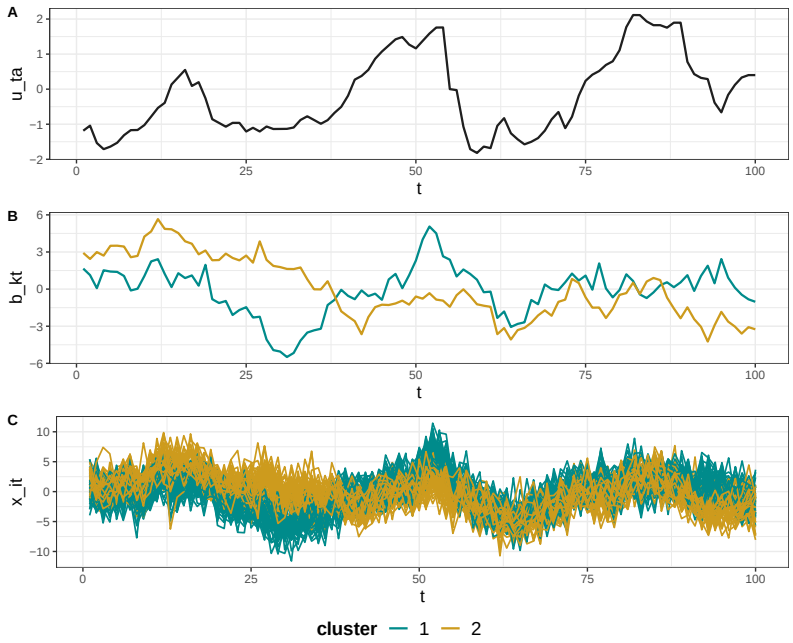

FIGURE 4.3 – Simulation d'un jeu de données de 100 observations de 100 pas de temps. (A) correspond à l'effet des facteurs exogènes, commun aux deux clusters. (B) représente les profils de classes simulés comme des processus autorégressifs. (C) représente l'ensemble des n observations simulées à l'aide des éléments précédents.

les classes d'appartenance de ces séquences et d'estimer les profils de classes (4.3(B)) et l'effet régressif (4.3(A)). La sous-section suivante propose un exemple d'estimation à partir du modèle.

4.4.2 Estimation des paramètres du modèle sur un jeu de données simulées

Itération et croissance de la borne inférieure

L'algorithme 1 est itératif et a pour objectif de mettre à jour les paramètres afin d'approximer le maximum de vraisemblance en maximisant la borne inférieure. À chaque itération, les paramètres sont mis à jour et la borne inférieure croît. Le graphique (4.4) représente l'erreur quadratique moyenne entre les centres de classes estimées et simulées (A), l'erreur quadratique moyenne entre l'effet régressif estimé et simulé (B) et la borne inférieure de la log-vraisemblance à chaque itération (C).

Classification

Tout d'abord, les classes d'appartenance des observations $\mathbf{x}_i$ sont estimées en utilisant la méthode du maximum *a posteriori* à partir des probabilités *a posteriori*. Ensuite, il est possible de comparer les classes estimées aux classes initialement simulées (figure 4.5).

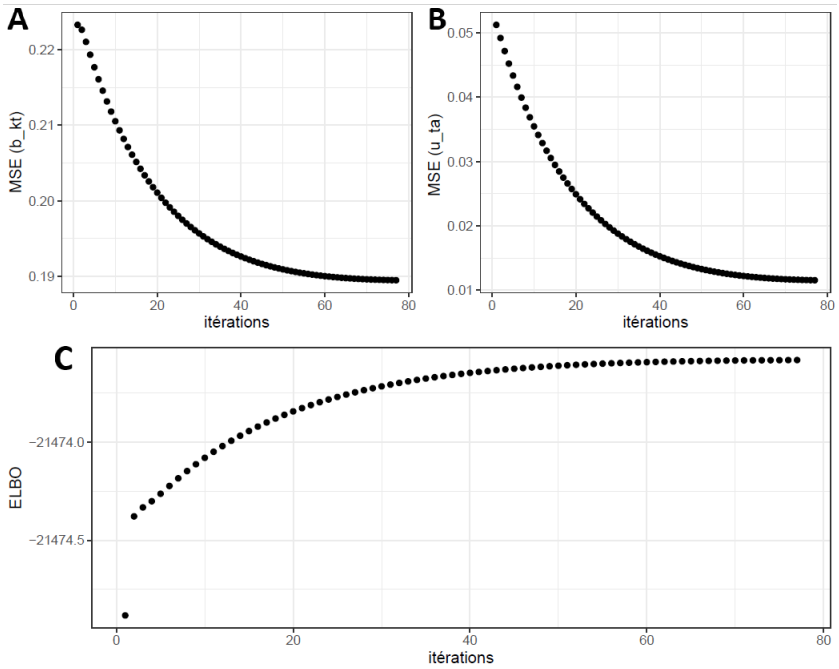

FIGURE 4.4 – Erreurs quadratiques moyennes et bornes inférieures de la vraisemblance calculées à chaque itération de l'algorithme d'estimation. (A) représente l'erreur quadratique moyenne calculée sur les centres de classes. (B) représente l'erreur quadratique moyenne calculée sur l'effet des facteurs exogènes. (C) représente la borne inférieure.

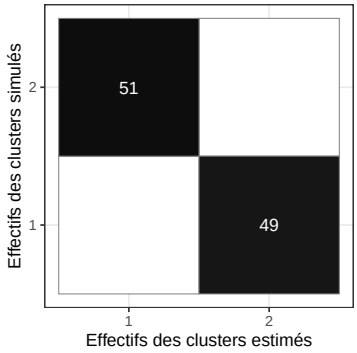

FIGURE 4.5 – Matrice de confusion entre les classes estimées et simulées.

Relabellisation des classes

Il est important de noter qu'il est possible que les labels des classes estimés ne correspondent pas à ceux des classes simulées. Cependant, afin de contrôler la capacité du modèle à bien estimer les profils de classes et les coefficients de régression, il faut associer chaque classe estimée à la classe simulée correspondante, dans la mesure du possible. Pour cela, avant de calculer les erreurs faites sur les centres de classes ou encore tracer les résultats sous la forme d'un graphique, une étape de relabellisation des classes est faite. Cette dernière consiste à associer les classes estimées et les classes simulées qui maximisent le taux de bonne classification. Dans le cas où la classification est parfaite, cette étape ne pose pas de soucis. Mais, lorsque des observations ne sont pas correctement classifiées, cette étape peut s'avérer

plus délicate.

Estimation des profils de classes et des coefficients de régression

Ensuite, le modèle permet d'estimer les profils de classes ainsi que les coefficients associés aux facteurs exogènes. L'estimation du modèle pour les données simulées présentées dans le graphique 4.3 permet de tracer les profils de classes estimés et simulés (4.6) et l'effet des facteurs exogènes estimé et simulé (4.7). On peut observer que l'estimation est plutôt satisfaisante et que l'erreur reste relativement faible. Afin de mieux évaluer ces performances, il est important de définir des critères précis dans la sous-section suivante.

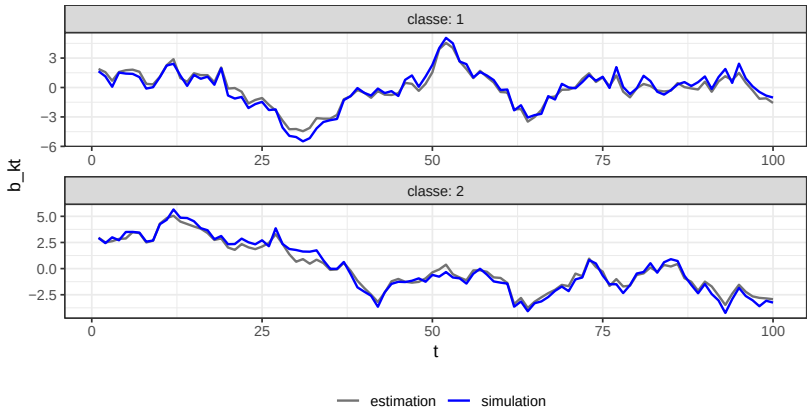

FIGURE 4.6 – Centres de classes estimés par le modèle (gris) et simulé (bleu).

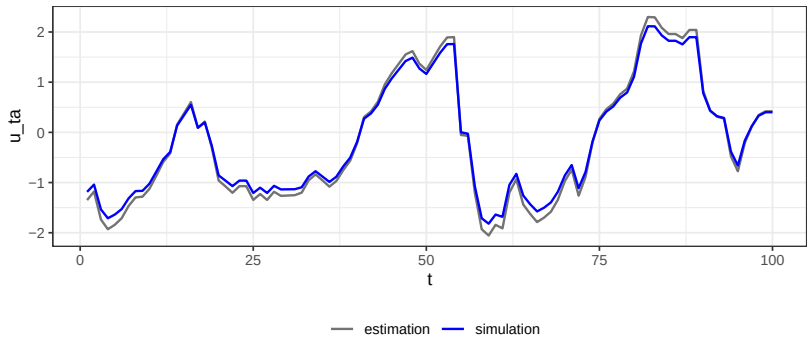

FIGURE 4.7 – Effet exogène estimé par le modèle (gris) et simulés (bleu).

4.4.3 Critères d'évaluation

Pour rappel, le modèle est supposé pouvoir identifier l'effet des variables exogènes, classifier les observations et estimer les centres de classe comme des processus dynamiques. L'objectif est d'évaluer le modèle sur ces trois aspects. Pour cette raison, les trois critères suivants ont été retenus.

La capacité du modèle à estimer les centres de classes est évaluée à partir de l'erreur quadratique moyenne calculée entre les centres de classes estimés et simulés :

$$\text{CRIT}_1 = \frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}(\hat{m}_{kt} - b_{kt})^2, \tag{4.32}$$

avec n, le nombre d'observations, T, la taille des séquences, K le nombre de clusters, et b_{kt} désignant les centres de clusters simulés et $\hat{m}_{kt}$ les centres de clusters estimés.

La capacité du modèle à estimer les effets des facteurs exogènes $\mathbf{u}_t$, qui désigne le vecteur de facteurs exogènes au temps t, avec $t = 1...T$, est évaluée à partir de l'erreur quadratique moyenne entre les effets régressifs estimés et simulés :

$$\text{CRIT}_2 = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{u}_t'\hat{\mathbf{a}} - \mathbf{u}_t'\mathbf{a})^2, \tag{4.33}$$

avec $\mathbf{a}$ et $\hat{\mathbf{a}}$ désignant respectivement les coefficients simulés et estimés.

La capacité du modèle à classifier les observations est évalué à partir du taux de bonne classification :

$$\text{CRIT}_3 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{z_i == \hat{z}_i\}}, \tag{4.34}$$

avec n, le nombre d'observations, et z_i le cluster de i et $\hat{z}_i$ le cluster estimé.

Ces critères seront utilisés pour tester les performances du modèle proposé et le comparer à deux autres modèles qui sont présentés dans la sous-section suivante.

4.4.4 Modèles de référence

Les trois critères d'évaluation des performances ont été calculés pour le modèle dynamique proposé puis comparés à deux modèles de base. L'objectif est de valider le modèle proposé et de confirmer que la méthode d'inférence variationnelle et les choix de modélisation sont appropriés pour les centres de clusters dynamiques. Pour mener à bien cette comparaison, il est utile de définir deux autres modèles de comparaison afin de justifier la modélisation dynamique des centres de classes et la méthode d'estimation choisie. Ces modèles ne sont pas des méthodes standards issues de la littérature, mais peuvent être définis comme des mélanges contraints de modèles de régression [Hurn et al., 2000; Desarbo and Wedel, 2002] où, à l'exception du biais, les coefficients sont communs aux clusters. Dans le premier modèle appelé "Centres constants", le biais ne dépend pas du temps, alors que dans le second modèle en dépend. Ce deuxième modèle est plus proche de celui proposé et constitue une bonne base de comparaison.

Premier modèle de référence : Régression linéaire et estimation de centres de classes constants ("Centres constants")

MODÉLISATION

On considère les données $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ avec $\mathbf{x}_i = (x_{it}, ..., x_{iT})$ et le modèle suivant :

$$x_{it} = b_{z_i} + \mathbf{u}_t'\mathbf{a} + \epsilon_{it}, \tag{4.35}$$

avec ϵ_{it} un bruit gaussien centré de variance $\sigma^2_{z_i}$ et z_i désigne la classe d'appartenance de l'observation i. Le modèle peut être considéré comme un mélange de régressions avec un terme de biais qui ne dépend pas du temps.

ESTIMATION

L'estimation de ce modèle sera réalisée en deux temps. D'abord, les coefficients régressifs sont estimés à partir de la formule suivante :

$$\hat{\mathbf{a}} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{u}_t\mathbf{u}_t'\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{u}_t x_{it}\right). \tag{4.36}$$

Ensuite, la classification et les centres de classes b_k seront estimés à partir des données $\tilde{x}_{it} = x_{it} - \mathbf{u}_t'\hat{\mathbf{a}}$. La difficulté réside dans le fait qu'on cherche une classe d'appartenance pour chaque séquence observée x_i alors que les centres de classes ne sont pas temporellement dépendants. Par conséquent, le choix a été fait d'estimer un modèle de mélange gaussien avec les paramètres (b_k, σ^2_k) et les proportions π_k à partir des données observées $\tilde{x}_{it}$. Cette estimation, via un algorithme EM, permet d'obtenir des probabilités *a posteriori* τ_{ikt}. Or, on se place dans un cadre où on cherche à classifier la séquence x_i et non chaque observation à chaque pas de temps. Par conséquent, une dernière étape de l'estimation consiste à déterminer $\hat{z}_i$ en fonction des probabilités *a posteriori* τ_{ikt} telle que :

$$\hat{z}_i = \underset{k}{\operatorname{argmax}}\ (\hat{\tau}_{ikt})_{k=1,..K}. \tag{4.37}$$

Ce modèle et l'estimation de ce dernier permettent de classifier les observations $\mathbf{x}_i$, d'estimer des centres de classes non-dépendants du temps et d'estimer l'effet régressif associé aux facteurs exogènes.

Second modèle de référence : Régression linéaire et classification K-means ("Régression + K-means")

MODÉLISATION

On considère les données $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ avec $\mathbf{x}_i = (x_{it}, ..., x_{iT})$. Le modèle est le suivant :

$$x_{it} = b_{z_i t} + \mathbf{u}_t'\mathbf{a} + \epsilon_{it}, \tag{4.38}$$

avec ϵ_{it} un bruit gaussien centré de variance $\sigma^2_{z_i}$ et z_i désigne la classe d'appartenance de l'observation i. Le modèle peut-être considéré comme un mélange de régressions avec un terme de biais qui dépend du temps.

ESTIMATION

L'estimation est réalisée en deux temps. D'abord, les coefficients régressifs sont estimés à partir de la formule (4.36). Ensuite, la classification et les centres de classes b_k seront estimés à partir des données retraitées $\tilde{x}_{it} = x_{it} - \mathbf{u}'_t\hat{\mathbf{a}}$. On utilise un algorithme K-means sur les données $\tilde{\mathbf{x}}_i$ afin d'identifier les classes d'appartenance de chaque observation $\mathbf{x}_i$ et les centres de classes b_{kt}.

4.4.5 Simulation de jeux de données variés

Afin de tester le modèle et ses limites, l'objectif est de générer divers ensembles de données pour s'assurer qu'un grand nombre de cas différents aient été testés. La figure suivante présente les différents cas qui seront utilisés pour générer des ensembles de données. Les jeux de données sont générés en jouant sur le niveau de difficulté du problème de classification et sur la taille des jeux de données. Le schéma 4.8 récapitule les différents cas de figure considérés pour simuler ces jeux de données.

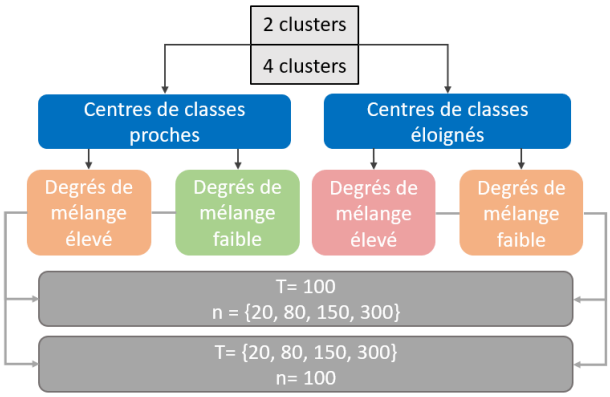

FIGURE 4.8 – Schéma explicatif des différents jeux de données simulées. Au total, pour 2 et 4 clusters, 8 cas de figure sont considérés selon l'écartement des centres de classes, le degré de mélange des classes et la taille des jeux de données.

Pour toutes ces configurations, des paramètres communs ont été utilisés pour les simulations. Le tableau 4.2 présente ces valeurs.

TABLE 4.2 – Paramètres utilisés lors de la simulation des jeux de données pour la comparaison des performances des modèles.

	(π_k)	(μ_{k0})	(σ^2_{k0})	(v^2_k)	(w^2_k)	(Φ_k)	**a**
K=2	(0.5, 0.5)	(6.5, 6.5)	(1, 1)	(2, 2)	(1, 1)	(0.9, 0.9)	(2.5, 1.8)
K=4	(0.25, 0.25, 0.25, 0.25)	(6.5, 6.5, 6.5, 6.5)	(1, 1, 1, 1)	(1.5, 2, 1.5, 2)	(1, 1, 1, 1)	(0.9, 0.9, 0.9, 0.9)	(2.5, 1.8)

À noter que les facteurs exogènes considérés sont les mêmes que ceux utilisés dans

l'exemple de la sous-section 4.4.1. De ce fait, les données de la Figure 4.2 représentent les vecteurs de facteurs exogènes des jeux de données simulées.

Différents niveaux de difficulté

Le premier élément qu'il est possible de faire varier pour générer des données variées est le niveau de difficulté du modèle. En effet, dans le cadre d'un problème de classification, on peut imaginer différents jeux de données pour lesquels les classes sont plus ou moins évidentes. En effet, en générant des données issues de classes plus ou moins distinctes ou avec des classes plus ou moins mélangées, il est possible de construire de jeux de données selon différents niveaux de difficulté.

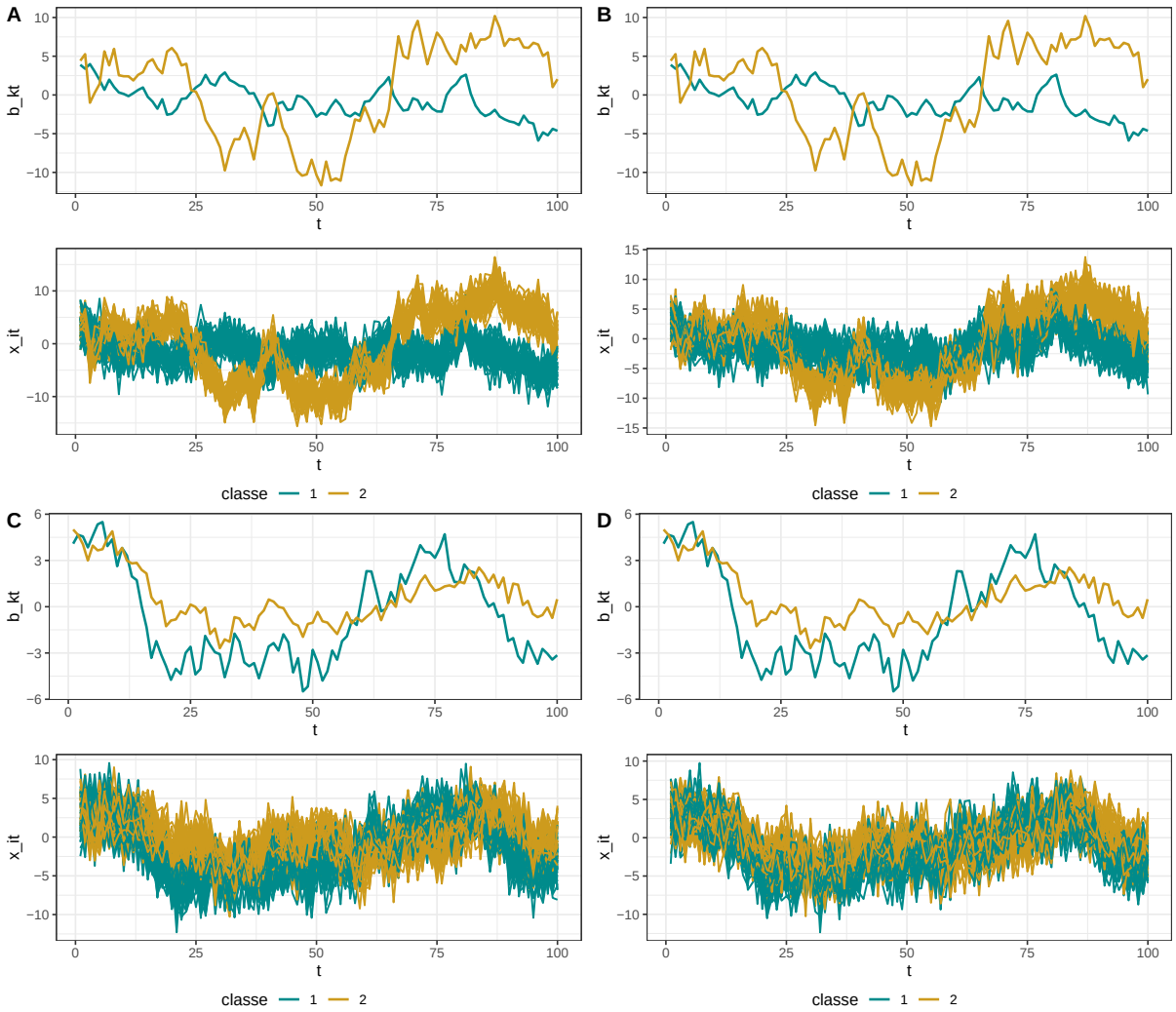

Figure 4.9 – Exemple de jeux de données, à deux classes, selon les niveaux de difficulté. (A) correspond au cas où les centres de classes sont distincts et le degré de mélange faible. (B) correspond au cas où les centres de classes sont distincts et le degré de mélange relativement plus grand. (C) correspond au cas où les centres de classes sont proches et le degré de mélange faible. (D) correspond au cas où les centres de classes sont proches et le degré de mélange relativement plus grand.

Le graphique 4.9 représente quatre jeux de données pour différents niveaux de difficulté. La différence entre les jeux de données (A, B) et (C, D) réside dans les centres de classes

générés. En effet, pour les deux premiers jeux de données, les centres de classes sont distincts, ils ont été générés de sorte à être assez éloignés les uns des autres. Tandis que pour les deux autres jeux de données (C, D), les centres de classes ont été générés de sorte à être rapprochés les uns des autres. Ensuite, la différence entre les jeux de données (A, C) et (B, D) réside dans le degré de mélange des classes. C'est-à-dire que, lorsqu'une séquence x_{it} est simulée, il est possible de considérer que l'observation i n'appartient qu'à une seule classe, ou il est possible de considérer que x_{it} est le résultat d'un mélange, plus ou moins hétérogène, entre plusieurs classes.

DIFFÉRENTES TAILLES DE JEUX DE DONNÉES

Le modèle est évalué en générant divers ensembles de données avec deux et quatre classes, et différents nombres d'observations. Pour chaque configuration, les modèles ont été testés sur deux cents ensembles de données différents. Tout d'abord, nous considérons la fenêtre temporelle fixe $T = 100$ et faisons varier le nombre d'observations ($n = 20$, $n = 80$, $n = 150$ et $n = 300$). Ensuite, nous fixons le nombre d'observations à $n = 100$ et fixons la fenêtre temporelle à $T = 20$, $T = 80$, $T = 150$ et $T = 300$. Cette stratégie permet d'analyser comment le modèle se comporte avec des jeux de données plus ou moins grands. De plus, dans le cas de données temporelles, on peut se demander si l'estimation est meilleure lorsque la séquence observée est plus longue.

4.4.6 Résultats obtenus

Le tableau 4.3 présente la valeur moyenne des critères pour les trois modèles dans différents cas. Pour rappel, le "modèle dynamique" correspond au modèle de classification proposé avec centres dynamiques et effet exogène commun. Le modèle "centres constants" correspond au premier modèle de référence. Il s'agit d'un modèle de classification avec des centres de classes constants. Le modèle "Régression + K-means" correspond au second modèle de référence et consiste en une estimation en deux temps avec d'abord l'estimation de l'effet régressif puis une classification K-means. Le modèle "Centres Constants" présente de moins bonnes performances sur la base des trois critères dans chacun des cas présentés.

La figure 4.10 montre les diagrammes en boîte des trois critères calculés pour le modèle dynamique à effet régressif commun et le modèle "Régression + K-means", sur 400 jeux de données avec différents niveaux de difficulté, et quatre clusters lorsque le nombre d'observations est égal à $(n = 20, T = 100)$, $(n = 80, T = 100)$ et $(n = 150, T = 100)$. Le modèle "Centres Constants", qui ne fournit pas de bonnes performances, n'est pas présenté pour des raisons de lisibilité. Pour rappel, les niveaux de difficulté dépendent de la similarité entre les profils de classes et du degré de mélange des celles-ci. Par exemple, des profils de classes similaires ou fortement mélangés correspondent au niveau de difficulté les plus élevés.

Ces résultats montrent différentes performances pour les deux modèles. En effet, le modèle dynamique proposé, pour les trois critères, semble être plus précis sur les jeux de données

TABLE 4.3 – Résultats de performances moyennes obtenus pour les trois modèles sur 400 jeux de données pour chaque taille d'échantillon considérée. Le CRIT_1 correspond à l'erreur quadratique moyenne calculée sur les centres de classe, CRIT_2 correspond à l'erreur quadratique moyenne calculée sur les effets exogènes et CRIT_3 correspond au taux de classification. Pour les trois critères, le modèle proposé fournit les meilleures performances. De plus, plus il y a d'observations, plus le modèle est précis.

		$CRIT_1$			$CRIT_2$			$CRIT_3$		
	T = 100	n = 20	n = 80	n = 150	n = 20	n = 80	n = 150	n = 20	n = 80	n = 150
K = 2	Centres Constants	7.181	7.176	7.213	0.311	0.309	0.307	0.595	0.601	0.599
	Régression + K-means	0.906	0.6	0.553	0.223	0.224	0.223	0.996	0.999	0.999
	Modèle dynamique	**0.684**	**0.465**	**0.446**	**0.156**	**0.161**	**0.189**	**0.996**	**0.999**	**0.999**
K = 4	Centres Constants	10.01	9.799	9.452	0.142	0.133	0.131	0.45	0.473	0.47
	Régression + K-means	1.778	1.028	0.95	0.142	0.133	0.131	0.925	0.989	0.993
	Modèle dynamique	**1.571**	**1.026**	**0.94**	**0.101**	**0.074**	**0.073**	**0.929**	**0.997**	**0.998**
	n=100	T=80	T = 150	T = 300	T=80	T = 150	T = 300	T=80	T = 150	T = 300
K = 2	Centres Constants	7.372	7.227	7.158	0.123	0.105	0.104	0.576	0.59	0.584
	Régression + K-means	1.457	1.457	1.46	0.123	0.105	0.104	1	1	1
	Modèle dynamique	**1.382**	**1.339**	**1.278**	**0.074**	**0.055**	**0.057**	**1**	**1**	**1**
K = 4	Centres Constants	6.573	6.438	6.369	0.066	0.057	0.058	0.632	0.712	0.768
	Régression + K-means	1.773	1.714	1.672	0.061	0.051	0.049	0.992	0.995	0.993
	Modèle dynamique	**1.746**	**1.671**	**1.569**	**0.04**	**0.026**	**0.024**	**0.996**	**0.999**	**1**

simulées. On constate que pour le taux de classification (Figure 4.10 (C)), le modèle "Régression + K-means" présente des valeurs plus extrêmes que le modèle dynamique pour $n = 150$ et $n = 80$ car ce dernier estime des probabilités d'appartenance et un mélange de clusters alors que le premier estime une classification stricte. Le modèle "Régression + K-means" a plus de difficulté à classifier des observations qui résultent d'un mélange de plusieurs clusters. En outre, nous pouvons noter que plus l'ensemble de données contient d'observations, plus le modèle dynamique est précis sur la base des trois critères.

La figure 4.11 montre les diagrammes en boîte des trois critères calculés pour deux des modèles, sur quatre cents ensembles de données avec différents niveaux de difficulté, avec quatre clusters lorsque le nombre d'observations est égal à ($n = 100, T = 80$), ($n = 100, T = 150$) et ($n = 100, T = 300$). Comme précédemment, le troisième modèle n'est pas représenté, mais les résultats sont affichés dans le tableau 4.3. Les diagrammes en boîte ont été obtenus en utilisant 400 ensembles de données pour chaque taille d'échantillon considérée avec différents niveaux de difficulté. Ces niveaux de difficulté sont gérés en utilisant la distance entre les profils de classes simulés et le degré de mélange des clusters.

Selon les trois critères, le modèle dynamique à effet régressif commun est plus performant que le modèle "Régression + K-means". Les moyennes du premier critère (Figure 4.11(A)) sont proches, mais le modèle dynamique est plus performant sur l'estimation de l'effet exogène et présente moins de valeurs extrêmes pour le taux de classification. Cela peut être dû aux cas difficiles de clusters fortement mélangés ou de centres de classe similaires. De plus, dans ce cas, plus la séquence T est longue, plus le modèle dynamique est précis.

Les résultats, présentés dans la Figure 4.11 et la Figure 4.10, montrent que les valeurs

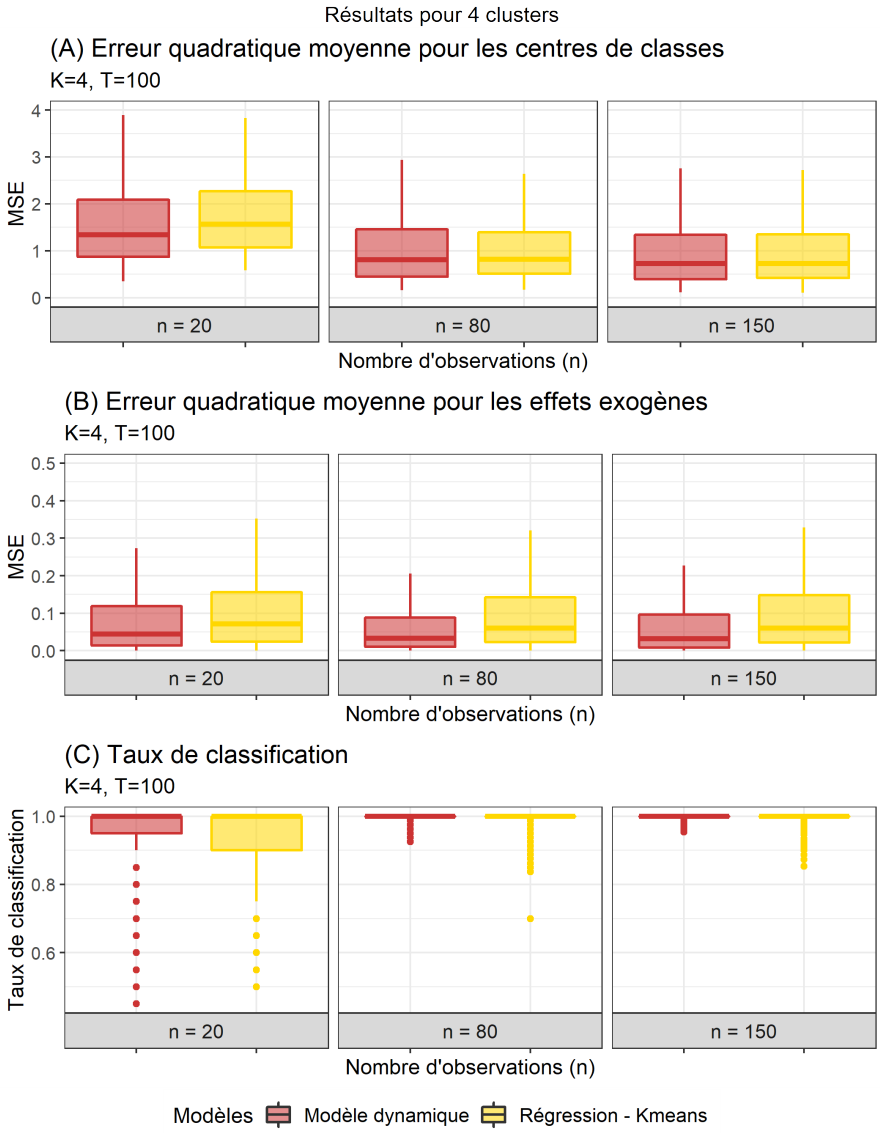

FIGURE 4.10 – Box-plot des trois critères obtenus pour le modèle complet proposé (rouge) et le modèle de régression en deux étapes (jaune) avec un ensemble de données de 100 pas de temps et différents nombres d'observations (n=20, n=80 et n=150) et deux clusters.

des deux premiers critères diminuent avec la taille de la fenêtre temporelle (T) et le nombre d'observations (n). Comme on peut s'y attendre, plus il y a de données, plus le modèle est précis. Les figures et les tableaux précédents montrent également que, sur les trois critères, le modèle dynamique à effet régressif commun est plus performant que les autres approches. Les performances du modèle dynamique par rapport au modèle "Centres constants" soulignent l'intérêt d'estimer les profils de classes de manière dynamique. En effet, ce dernier modèle est basé sur l'hypothèse de centres de classe constants dans le temps et classifie les séries temporelles en utilisant le clustering réalisé pour chaque pas de temps. Ces résultats montrent que ce modèle n'est pas bien adapté aux données simulées. Il est cependant important de noter que l'étape de régression, utilisée pour les deux modèles de référence, fournit une estimation relativement bonne de l'effet exogène.

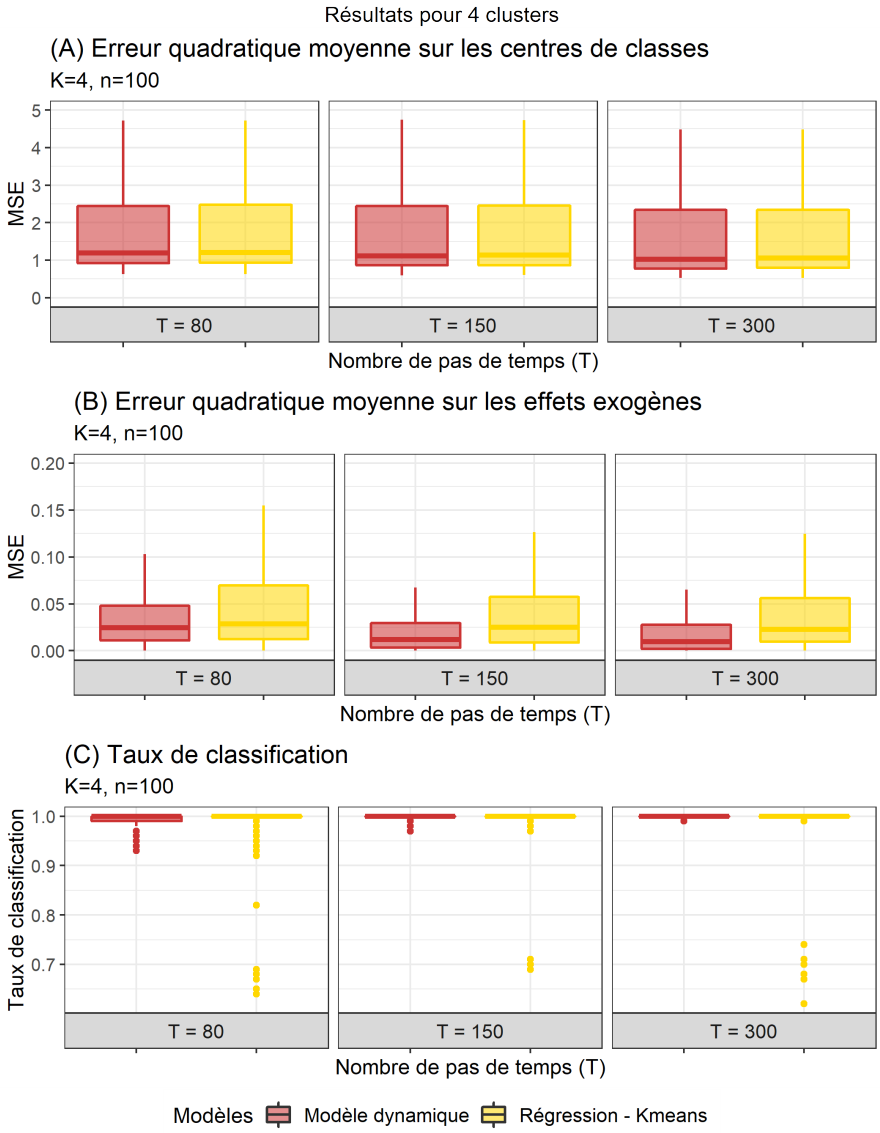

FIGURE 4.11 – Box-plot des trois critères obtenus pour le modèle complet proposé (rouge) et le modèle de régression en deux étapes (jaune) avec un ensemble de données de 100 observations et différents nombres de pas de temps (T=80, T=150 et T=300) et quatre clusters.

Dans ce premier modèle, on a fait également l'hypothèse que l'effet régressif des facteurs exogènes était commun à l'ensemble des observations. On peut alors supposer, comme c'est le cas dans les modèles de mélange de régressions [Hurn et al., 2000], que les effets régressifs sont spécifiques à chaque cluster. La section suivante est dédiée à présenter l'extension du précédent modèle avec pour particularité d'estimer des effets exogènes propres à chaque classe.

4.5 Extension du modèle de classification proposé avec estimation d'effets exogènes spécifiques à chaque cluster

Les modèles de mélange de régressions permettent d'estimer des effets régressifs propres à chaque cluster. Cependant, dans le cadre des données sujettes à la fois à des effets exogènes connus et à des effets endogènes inconnus, le mélange de régressions pourrait ne pas suffire. Dans cette section, on propose une forme moins contrainte du précédent modèle où les coefficients régressifs sont propres à chaque classe.

4.5.1 Intérêt et motivation

Jusqu'alors, on a fait l'hypothèse que l'ensemble des logements d'une zone ou d'un immeuble sont influencés de la même manière par des facteurs exogènes connus et mesurables comme des données météorologiques ou calendaires. Or, l'isolation, l'exposition ou encore la taille des ouvertures sur l'extérieur peuvent avoir un impact sur l'effet que ces facteurs météorologiques ont sur les variables mesurées au sein de ces logements. Par conséquent, il semble intéressant d'estimer ces effets au sein de chaque classe.

Le principe d'estimer des effets régressifs au sein de chaque cluster dans le cadre d'un modèle de classification est l'idée centrale des modèles de mélange de régressions. Mais ces modèles nécessitent une forme explicite des facteurs exogènes. Dans notre cas, ayant fait l'hypothèse qu'une partie seulement des données est expliquée par des facteurs connus, le mélange de régressions peut sembler limité pour modéliser de telles données. Le modèle proposé par la suite permet à la fois de classifier les observations, d'estimer des coefficients régressifs pour chaque cluster et d'estimer des profils de classes dynamiques relatifs à des effets endogènes non explicables par des variables connues et observables.

Cette section est destinée à présenter ce modèle et l'algorithme d'estimation utilisé pour l'estimer. Ensuite, les performances de ce modèle seront évaluées selon les critères d'évaluation présentés dans la section 4.4.3 et comparées celles issues d'un modèle de référence qui sera introduit par la suite.

4.5.2 Modélisation

En reprenant les mêmes notations que dans la section 4.2, le modèle s'écrit :

$$\forall i \in [\![1;n]\!],\ \ \forall t \in [\![1,T]\!];\quad x_{it} = \sum_{k=1}^{K} z_{ik}\left(\mathbf{u}_t'\mathbf{a}_k + b_{kt}\right) + e_{it}, \tag{4.39}$$

où z_{ik} est une variable muette égale à 1 si l'observation i appartient à la classe k et 0 sinon. De plus, le processus $(b_{kt})_{t=1,\dots,T}$ correspond au centre de la classe k, e_{it} est un terme de bruit et $\mathbf{a}_k \in \mathbb{R}^p$ désigne les coefficients associés aux facteurs exogènes pour chaque cluster.

Comme pour le précédent modèle de classification dynamique avec des effets exogènes communs, l'un des objectifs est de prendre en compte et de modéliser la dynamique des profils de classes. Dans ce cas aussi, les clusters sont caractérisés par des processus autorégressifs de premier ordre :

$$\forall t \in [\![1,T]\!]; \forall k \in [\![1,K]\!]; \quad b_{kt} = \Phi_k b_{kt-1} + \nu_{kt} \tag{4.40}$$

Identification du modèle et contraintes

Le modèle défini par l'équation (4.39) n'est pas identifiable. Afin d'assurer son identifiabilité, le centrage des profils latents est effectué, pour chaque classe, tel que : $\forall k,\ \sum_{t=1}^{T} b_{kt} = 0$.

4.5.3 Estimation

L'estimation de ce modèle est semblable à l'estimation du précédent modèle pour lequel les effets régressifs sont communs. En reprenant la méthode d'inférence variationnelle présentée dans le chapitre 3, on définit la borne inférieure de la log-vraisemblance suivante :

$$\begin{aligned}
F(\mathbf{m},\boldsymbol{\tau},\boldsymbol{\lambda},\boldsymbol{\Theta}) &= \mathbf{E}_q(\mathscr{L}_c(\boldsymbol{\Theta})) + H(q) \\
&= \sum_{i,t,k} \tau_{ik} \left(\log(\pi_k \varphi(x_{it}; m_{kt} + \mathbf{u}_t' \mathbf{a}_k, v_k^2)) - \frac{1}{2}\lambda_k(v_k^{-2}) \right) + \sum_{k,t} \log\left(\varphi(m_{kt}; \Phi_k m_{kt-1}, w_k^2)\right) \\
&- \frac{1}{2}\lambda_k((w_k^{-2}) + (w_k^{-2}\Phi_k^2)) + \sum_k \log\left(\varphi(m_{k0}, \mu_{k0}, \sigma_{k0}^2)\right) - \frac{1}{2}\lambda_k(\sigma_{k0}^{-2}) \\
&- \sum_{i,k} \tau_{ik}\log(\tau_{ik}) + \frac{d(T+1)}{2} \sum_k \log(2\pi e) + \log(\lambda_k).
\end{aligned} \tag{4.41}$$

L'estimation des paramètres repose, ici aussi, sur un algorithme itératif de mise à jour des paramètres et paramètres variationnels dont les formules sont obtenues par maximisation de la borne inférieure (4.41).

4.5.4 Évaluation des performances sur des données simulées

Ce modèle doit être validé sur des données simulées et contrôlées afin de s'assurer de sa capacité à classifier et à estimer les effets régressifs et les profils de classes. Cette sous-section est dédiée à l'évaluation des performances de ce modèle. Tout d'abord, la méthode de simulation des données sera présentée, ensuite les critères d'évaluation seront rappelés. Afin d'évaluer les performances du modèle, deux modèles de comparaison sont introduits pour servir de base de comparaison. Pour finir, les deux modèles dynamiques proposés (avec des effets régressifs communs et avec des effets régressifs propres aux classes) seront comparés sur différents jeux de données.

Simulation des données

La simulation d'un ensemble de données pour ce modèle se déroule de la même manière que dans la section 4.4.1 à la différence que les effets exogènes sont définis pour chaque cluster. C'est-à-dire qu'au lieu de définir un vecteur de coefficient **a**, il faut définir K vecteurs de coefficients pour chaque cluster simulé.

TABLE 4.4 – Paramètres utilisés pour la simulation du jeu de données.

	(π_k)	(μ_{k0})	(σ^2_{k0})	(v^2_k)	(w^2_k)	(Φ_k)	**a**
K=2	(0.5,0.5)	(4.5, 6)	(1, 1)	(2, 2)	(1, 1)	(0.99, 0.95)	(2.5, 1.8)

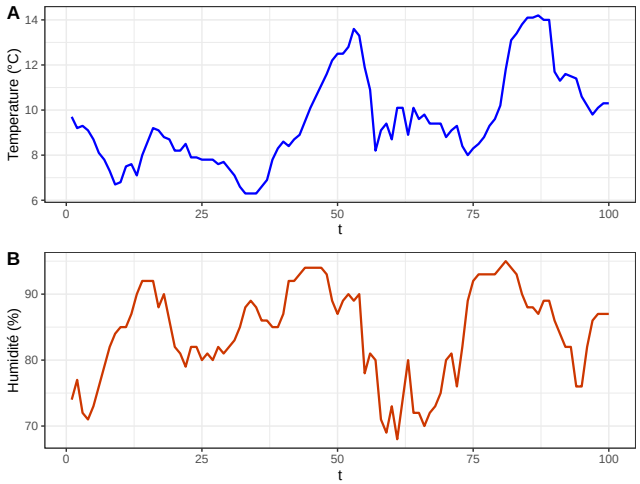

FIGURE 4.12 – Données météorologiques utilisées comme facteurs exogènes lors de la simulation des données. À noter que ces données ont été centrées et réduites pour la simulation.

Les données d'entrée, qu'on appelle "observations", sont les séquences $\mathbf{x}_i$ généréesà partir des paramètres présentés dans le tableau 4.4. Ces simulations sont tracées sur le graphique 4.13(C). L'objectif du modèle est de déterminer les classes d'appartenance de ces séquences et d'estimer les profils de classes (4.13(B)) ainsi que les effets régressifs (4.13(A)). La sous-section suivante propose un exemple d'estimation à partir du modèle.

Critères d'évaluation

Comme précédemment, on définit 3 critères d'évaluation et de comparaison des performances du modèle proposé.

- L'erreur quadratique moyenne entre les centres de classes estimés et simulés.
- L'erreur quadratique moyenne entre les effets exogènes estimés et simulés.
- Le taux de classification entre les classes d'appartenance estimées et simulées.

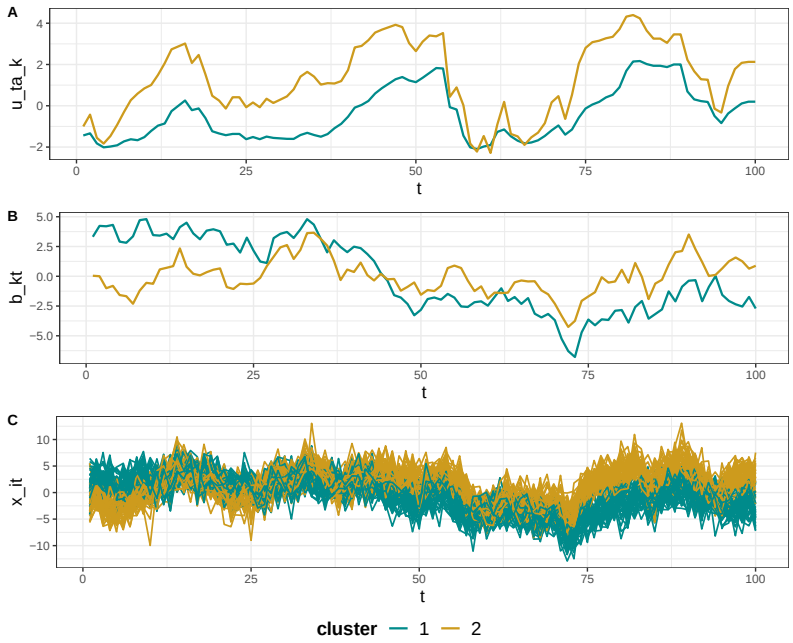

FIGURE 4.13 – Simulation d'un jeu de données de 100 observations de 100 pas de temps. (A) correspond aux effets des facteurs exogènes pour les deux clusters. (B) représente les profils de classes simulés comme des processus autorégressifs. (C) représente l'ensemble des n observations simulées à l'aide des éléments précédents.

Modèle de référence : K-MEANS + REGRESSION

L'intérêt de définir un modèle de référence est de pouvoir comparer les performances du modèle dynamique proposé sur des jeux de données identiques avec des modèles permettant de réaliser des tâches de classification et d'estimation similaires.

Le modèle de comparaison est un modèle simple, qui peut être considéré comme un modèle de mélange de régressions. On considère les données $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ avec $\mathbf{x}_i = (x_{it}, ..., x_{iT})$ et le modèle suivant :

$$x_{it} = \mathbf{u}_t' \mathbf{a}_{z_i} + b_{z_i t} \epsilon_{it}, \tag{4.42}$$

avec ϵ_{it} un bruit gaussien centré de variance $\sigma^2_{z_i}$ et z_i, la classe d'appartenance de l'observation i.

L'estimation des classes, centres de classes et effets exogènes repose sur une méthode en deux temps. D'abord, on utilise l'algorithme K-means pour la classification des données $(\mathbf{x}_1, ..., \mathbf{x}_n)$ afin de définir les classes d'appartenance des observations $(\hat{z}_i)$. Ensuite, dans un second temps, au sein de chaque classe, les effets exogènes sont estimés à partir de la formule suivante :

$$\hat{\mathbf{a}}_k = \left(n_k \sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \right)^{-1} \left(\sum_{z_i = k} \sum_{t=1}^{T} \mathbf{u}_t x_{it} \right). \tag{4.43}$$

Pour finir, les centres de classes b_{kt} sont estimés à partir des résidus issus de la régression

linéaire. En effet, on calcule les résidus moyens pour estimer les centres de classes tels que :

$$\forall t,\ \ \hat{b}_{kt} = \frac{1}{n_k} \sum_{z_i=k} (x_{it} - \mathbf{u}_t' \hat{\mathbf{a}}_k), \tag{4.44}$$

avec n_k, l'effectif du cluster k. Cette méthode permet d'estimer les coefficients, les profils de classes et les classes d'appartenance des observations et ainsi calculer les critères pour la comparaison.

Simulation de jeux de données variés

Comme pour le premier modèle, l'objectif est d'évaluer les performances du modèle sur un grand nombre de jeux de données variées. Pour cela, on utilise la même stratégie précédemment (voir schéma 4.8) avec la méthode de simulation présentée et avec les valeurs de paramètres présentées dans le tableau 4.5.

TABLE 4.5 – Paramètres utilisés lors de la simulation des jeux de données pour l'évaluation des performances du modèle dynamique avec estimation des coefficients propres à chaque classe. Deux facteurs exogènes sont considérés.

	(π_k)	(μ_{k0})	(σ^2_{k0})	(ν^2_k)	(w^2_k)	(Φ_k)	$\mathbf{a}_k$
K=2	$\pi_k = 0.5$	(6.5, 6.5)	(1, 1)	(2, 2)	(1, 1)	(0.9, 0.9)	[(3, 0.4) (0.9, 3)]
K=4	$\pi_k = 0.25$	$\mu_{k0} = 6.5$	$\sigma^2_{k0} = 1$	(1.5, 2, 1.5, 2)	$w^2_k = 1$	$\phi_k = 0.9$	[(4, 0.4), (0.9, 3), (2, 2), (4, 2.5)]

Il aurait été possible ici d'ajouter des niveaux de difficulté en jouant sur la valeur des coefficients avec des valeurs plus ou moins différentes d'un cluster à l'autre. Mais nous avons fait le choix de ne pas utiliser ce levier pour ne pas multiplier les cas à traiter. De plus, la sous-section suivante, destinée à la comparaison entre les deux modèles proposés, permet d'aborder les cas où les coefficients sont identiques pour l'ensemble des classes mais, sont estimés tout de même pour chaque classe.

4.5.5 Résultats de l'évaluation des performances du modèle proposé et comparaison avec le modèle de référence

La figure 4.15 montre les diagrammes en boîte des trois critères calculés pour le modèle dynamique avec estimation des effets exogènes pour chaque cluster et le modèle "K-means + Régression" avec deux clusters lorsque le nombre d'observations est égal à ($n = 20$, $T = 100$) et ($n = 80$, $T = 100$). Ces résultats montrent des performances différentes entre les deux modèles. Le modèle dynamique proposé, pour les trois critères, semble être plus précis sur les jeux de données simulées. La figure 4.14 représente les diagrammes en boîte des trois critères calculés pour deux des modèles, sur deux cents ensembles de données avec différents niveaux de difficulté, avec quatre clusters lorsque le nombre d'observations est égal à ($n = 100$, $T = 80$) et ($n = 100$, $T = 150$). Les diagrammes en boîte ont été obtenus en utilisant 400 ensembles de

données pour chaque taille d'échantillon considérée avec différents niveaux de difficulté. Ces niveaux de difficulté sont gérés en utilisant la distance entre les profils de classes simulés et le degré de mélange des clusters.

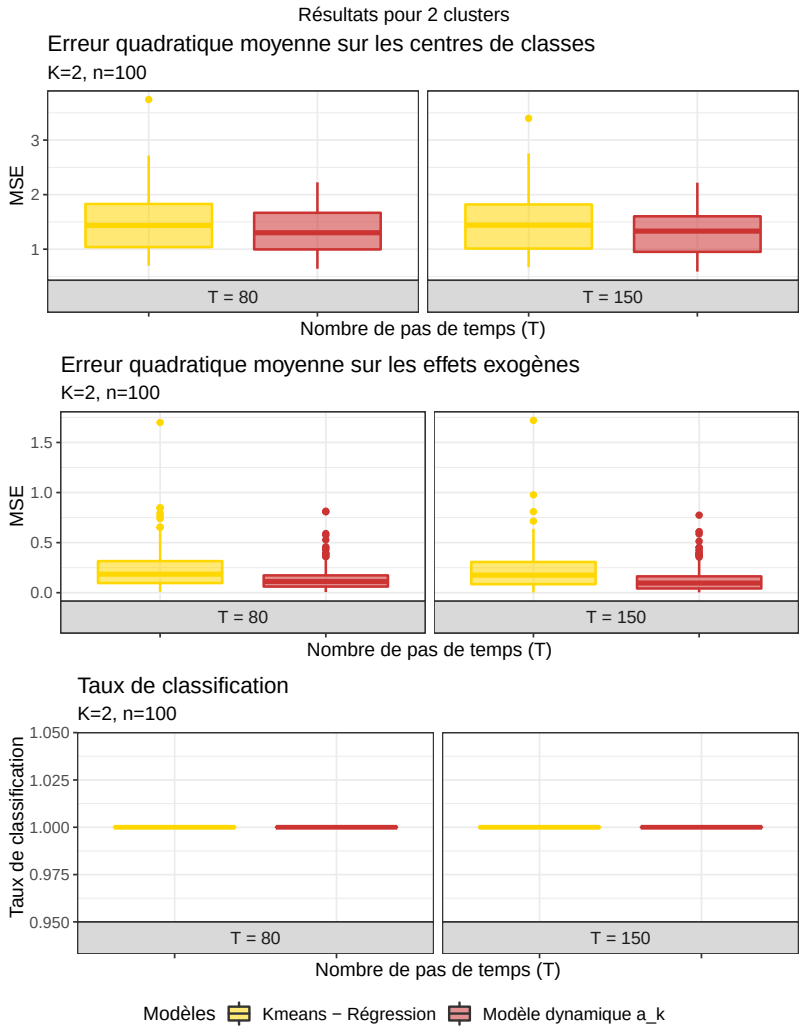

FIGURE 4.14 – Box-plot des trois critères obtenus pour le modèle complet proposé avec estimation des effets exogènes par classes (rouge) et le modèle de régression en deux étapes (K-means + Régression) (jaune) avec un ensemble de données de 100 observations et différents nombres de pas de temps (T=80, T=150) et deux clusters (K=2).

Les résultats présentés dans la Figure 4.15 montrent que les valeurs des deux premiers critères diminuent avec la taille de la fenêtre temporelle (T) et le nombre d'observations (n). Comme attendu, plus il y a de données, plus le modèle est précis. Les figures précédentes montrent également que, sur les trois critères, le modèle dynamique estimé avec un algorithme itératif est plus performant qu'une approche en deux temps basée sur l'algorithme K-means.

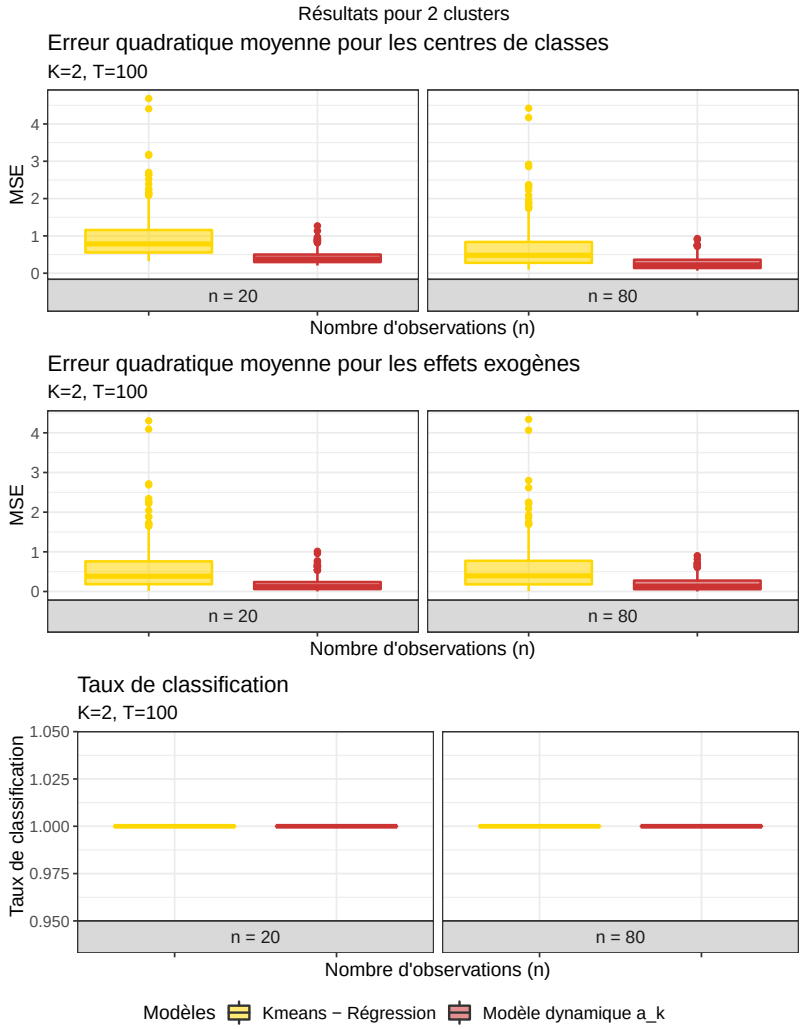

FIGURE 4.15 – Box-plot des trois critères obtenus pour le modèle complet proposé avec estimation des effets exogènes par classes (rouge) et le modèle de régression en deux étapes (jaune) avec un ensemble de données de 100 pas de temps et différents nombres d'observations (n=20, n=80) et deux clusters (K=2).

TABLE 4.6 – Valeurs moyennes des critères calculés pour 400 ensembles de données de niveaux de difficulté différents.

		$CRIT_1$		$CRIT_2$		$CRIT_3$	
T=100		n=20	n=80	n=20	n=80	n=20	n=80
K=2	K-means + régression	0.984	0.692	0.591	0.593	1	1
	Modèle dynamique avec effet régressif par classe	**0.426**	**0.27**	**0.185**	**0.187**	**1**	**1**
n=100		T= 80	T=150	T= 80	T=150	T= 80	T=150
K=2	K-means + régression	1.338	1.33	0.231	0.213	1	1
	Modèle dynamique avec effet régressif par classe	**1.217**	**1.188**	**0.138**	**0.132**	**1**	**1**

4.5.6 Comparaison des deux modèles sur des données simulées

Cette sous-section est dédiée à la comparaison et à l'analyse des deux modèles principaux présentés dans cette section. Pour rappel, les deux modèles d'intérêt sont des modèles de

classification pour des données temporelles dont les profils de classes sont estimés comme des processus autorégressifs et qui sont sujettes à des effets exogènes. Pour le premier modèle ("modèle **a**"), ces effets exogènes sont communs à l'ensemble des observations quelle que soit la classe d'appartenance, tandis que pour le second modèle ("modèle $\mathbf{a}_k$") les effets sont spécifiques à chaque classe.

L'avantage du premier modèle est que la séparation des effets est plus marquée et permet d'isoler un effet régressif commun qui est lié à la météo ou à d'autres données calendaires. Les effets endogènes estimés par les profils de classes sont spécifiques aux clusters et reflètent des habitudes ou des patterns de présence dans le cas de données d'appartement. Le second modèle a pour intérêt de mettre en évidence des différences de sensibilité à des facteurs exogènes. La séparation des effets, au sein d'un cluster, est alors moins marquée et le modèle a tendance à être moins performant pour estimer les centres de classes et les effets exogènes.

Méthode de comparaison

Afin de comparer ces deux modèles, on va se baser sur les critères présentés dans les sous-sections précédentes 4.4.3 et 4.5.4. Ces critères seront calculés sur des jeux de données, simulés à partir du premier modèle et du second modèle. Cela permet d'évaluer les performances du premier modèle sur des données pour lesquelles les effets exogènes sont différents d'une classe à l'autre, ainsi que d'évaluer le second modèle sur des données dont les effets régressifs sont communs.

Cinquante jeux de données ont été simulés à partir du premier modèle comme dans la figure 4.3, avec deux et quatre clusters, et la même chose à partir du second modèle, comme dans la figure 4.13. La comparaison se base sur 200 jeux de données.

Résultats de la comparaison

Le graphique ci-dessous (4.16) représente les diagrammes en boîtes des trois critères d'évaluation calculés pour les ensembles de jeux de données simulées via le premier modèle dynamique à effet régressif commun et via le second à effets régressifs par clusters. Les deux modèles ont été appliqués à ces deux ensembles de jeux de données avec deux clusters.

Ces résultats de comparaisons montrent deux choses importantes. La première est que, lorsque les effets sont communs à l'ensemble des classes, le modèle qui estime des effets pour chaque cluster est légèrement moins performant que l'autre. Ce résultat est dû au fait que la séparation des effets dans le modèle $\mathbf{a}_k$ est moins évidente. Malgré les contraintes d'identification ajoutées pour l'estimation, il est possible qu'une partie des effets exogènes soient estimés avec les profils de classes. La deuxième chose qu'on observe dans les graphiques 4.16 est que lorsque les données sont sujettes à des effets régressifs propres à chaque classe, alors le modèle $\mathbf{a}_k$ performe mieux que le modèle **a**. On peut également ajouter que l'écart de performance est bien plus important dans le cas de ces jeux de données que dans l'autre.

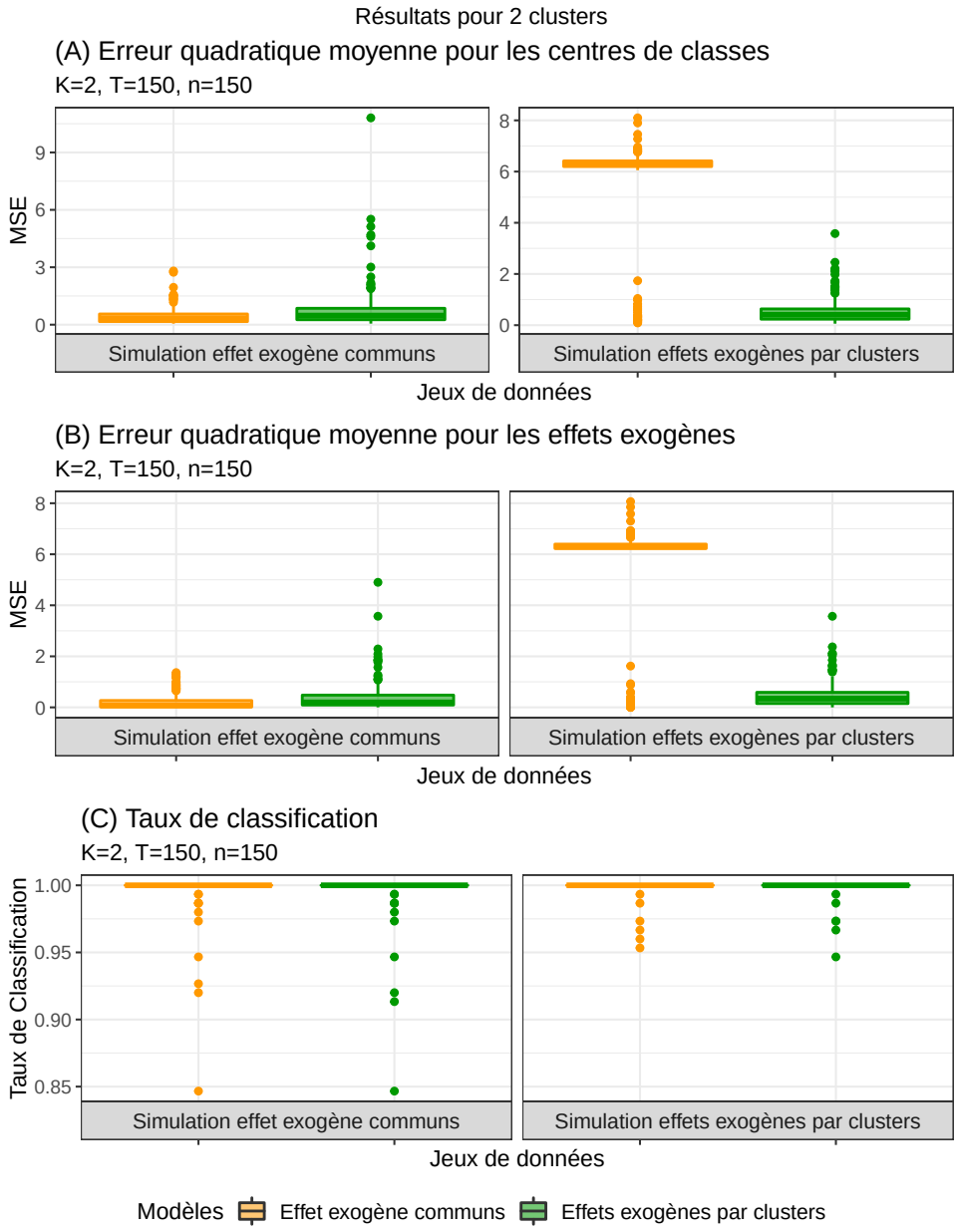

FIGURE 4.16 – Box-plot des trois critères obtenus pour le modèle dynamique avec estimation des effets exogènes communs (orange) et le modèle dynamique avec estimation des effets exogènes pour chaque cluster (vert) sur un ensemble de jeux de données simulées selon les deux modèles, avec $K = 2$, $T = 150$ et $n = 150$. Les diagrammes en boîte ont été obtenus en utilisant 100 jeux de données pour chaque cas considérés.

De ce fait, dans un contexte applicatif, si aucune information sur les potentielles différences de sensibilités à des facteurs exogènes n'est disponible, il serait plus intéressant d'utiliser le modèle $\mathbf{a}_k$. Par contre, dans le cas de données réelles ou simulées pour lesquelles il est certain ou très probable que les effets exogènes soient communs, il sera préférable d'utiliser le modèle $\mathbf{a}$.

4.6 Conclusion du chapitre

Ce chapitre a présenté dans un premier temps un modèle de classification de données temporelles qui permet d'estimer des effets régressifs communs et des profils de classes modélisés comme des processus autorégressifs afin de prendre en compte les dynamiques temporelles. Ce modèle est estimé à l'aide d'un algorithme basé sur l'inférence variationnelle qui permet d'approcher le maximum de vraisemblance via une procédure itérative. Cette méthode dynamique est évaluée et comparée sur la base de trois critères calculés à partir de jeux de données simulées et donc contrôlées. En effet, en simulant des données à partir du modèle, il est alors possible de calculer les erreurs d'estimation sur les effets régressifs, la classification et les profils de classes. Ces performances sont comparées avec deux modèles de référence. Ces modèles de référence ont été définis afin de constituer des bases de référence et de montrer l'intérêt d'estimer les dynamiques des centres de classe à l'aide de l'algorithme itératif proposé.

Ensuite, une extension du premier modèle dynamique est proposée. Ce second modèle permet d'estimer les effets régressifs au sein de chaque cluster. Ce modèle est également estimé à partir d'un algorithme d'inférence variationnelle. La comparaison est d'abord menée à partir d'un modèle de référence et de jeux de données simulées. Pour finir, les deux modèles dynamiques sont comparés sur des jeux de données simulées d'une part avec des effets régressifs communs et, d'autre part, avec des effets régressifs spécifiques à chaque classe. On tire de ces observations la conclusion que dans le cas où on ne possède aucune information *a priori* sur les effets exogènes, le modèle dynamique à effet commun semble moins indiqué que le second.

Pour conclure, ce chapitre a permis d'introduire, d'un point de vue méthodologique, deux modèles de classification avec centres de classes dynamiques et effets régressifs. La motivation sous-jacente à la construction de ces modèles réside dans l'analyse et dans la classification de comportements d'occupants de logements. En effet, on s'intéresse aux comportements de chauffage et de présence au sein de logements et le modèle a été construit dans le but d'identifier des patterns de comportements à partir de variables mesurées à l'intérieur d'un logement. Le prochain chapitre est dédié à l'application de ce modèle sur des données simulées à partir d'un modèle thermique afin de s'approcher d'un contexte réel et sur des données réelles de température au sein de maisons individuelles.

Chapitre 5

Application à des données thermiques simulées et réelles

Contents

5.1 Introduction

Les modèles de classification présentés dans le chapitre précédent visent à classifier des données tout en isolant les effets régressifs liés à des facteurs exogènes mesurables et à mettre en évidence les profils dynamiques des classes. Dans le contexte énergétique de la thèse, l'objectif est d'identifier, à partir des mesures de température intérieures, des classes de logements et les caractériser via des profils dynamiques. Le postulat de départ est de considérer que la température intérieure d'un logement est influencée, d'une part, par des facteurs exogènes tels que les conditions météorologiques extérieures ou encore des variables calendaires (saison, jour...) et d'autre part, par des comportements endogènes liés à l'activité, la présence, les habitudes de chauffage et les habitudes de vie des habitants.

Ce chapitre présente, dans un premier temps, l'application du modèle de classification avec centres de classes dynamiques et effets régressifs communs sur des données de température simulées via un modèle thermique. La méthode de simulation des données est détaillée, puis les résultats obtenus sur l'ensemble des appartements sont présentés. Ensuite, on cherchera à explorer la séparation des effets endogènes et exogènes sur un seul logement en laissant de côté l'aspect classification du modèle. Dans une seconde partie, le modèle est appliqué à un jeu de données réelles de la base de données REFIT [Firth et al., 2017] présentée dans le chapitre 2 et qui regroupe des données issues de maisons individuelles situées en Angleterre. Cette application permet de mettre en évidence un lien important entre les profils de classes et la présence moyenne des occupants au sein des pièces de vie. Pour finir, une application sur des données de parties communes issues de la base de données ANDRE, présentée dans le chapitre 2, est proposée. La classification des données de capteurs permet de mettre en évidence des groupes de capteurs qui se distinguent par leur sensibilité aux températures extérieures.

5.2 Application à des données simulées à l'aide d'un modèle thermique

Dans le chapitre 4, les modèles de classification dynamiques proposés ont été testés et comparés sur des données simulées statistiquement. Cependant, la motivation sous-jacente est l'application à des données thermiques pour estimer des comportements de chauffage, d'occupation type au sein de logements. L'acquisition de telles données est longue, coûteuse et peut poser des problèmes de protection de la vie privée. De plus, dans un souci d'évaluation des modèles, il est intéressant d'utiliser des données thermiques contrôlées. En effet, la simulation des données permet de connaître les facteurs ayant un effet exogène ainsi que les classes inobservées. Par conséquent, l'utilisation d'un modèle thermique du bâtiment pour la simulation d'un jeu de données paraît adaptée.

Ainsi, l'objectif de ce chapitre est de simuler un ensemble de données de température issues de plusieurs logements supposés appartenir au même immeuble. L'intérêt est de pouvoir considérer des données contextuelles communes à l'ensemble des logements afin d'identifier un effet exogène partagé. De plus, dans une optique de meilleure compréhension des comportements dans le cadre de la simulation de la consommation d'énergie à l'échelle d'un bâtiment, il est cohérent de s'intéresser à un ensemble de logements d'un même immeuble.

Dans un premier temps, le modèle thermique "monozone" ainsi que la démarche suivie pour la simulation des données seront présentés. Ensuite, le modèle de classification avec un effet exogène commun est appliqué aux données simulées et les résultats sont analysés. Dans un second temps, on s'intéresse à la capacité du modèle à séparer les effets exogènes des effets endogènes en laissant de côté l'aspect classification et en n'utilisant les données que d'un appartement à la fois. L'intérêt de cette dernière utilisation du modèle proposé est d'explorer la capacité du modèle à capter des effets liés à des variables calendaires et météorologiques, et à estimer un profil de classes dynamique lié à des facteurs endogènes non-observables.

5.2.1 La simulation thermique des données

Contexte et modèle thermique utilisé

La simulation des données thermiques permet de simuler des mesures réalistes de températures intérieures issues d'un ensemble de logements ayant des comportements de chauffage et d'occupation variés et pouvant être regroupés en classes homogènes.

Pour cela, on utilise un modèle de simulation thermique dont les données d'entrée et de sortie sont détaillées dans le schéma de la figure 5.1. Le modèle de simulation, décrit et utilisé dans [Djatouti et al., 2020, 2021], a été développé par des experts en thermique du bâtiment. Il s'agit d'un modèle thermique dit "monozone", qui permet de modéliser le comportement d'un appartement constitué d'une pièce. Ce modèle suppose qu'il n'y a que des échanges entre l'appartement en question et l'extérieur. Ainsi, on fait l'hypothèse qu'aucun échange thermique n'est possible entre les appartements d'un même immeuble.

Avec l'ensemble des entrées du modèle de simulation thermique défini, on obtient les données de température intérieure, de température de la paroi du mur en contact avec l'extérieur et la puissance des radiateurs associée via la résolution d'un système d'équations différentielles. Ce modèle est décrit plus en détail dans l'annexe A.

La démarche utilisée pour simuler des données de température intérieure d'un ensemble d'appartements

La génération des signaux de thermostats pour l'ensemble des appartements se déroule en 4 étapes qui seront détaillées par la suite. Dans un premier temps, il faut définir les caractéristiques des appartements et les conditions météorologiques extérieures. Ensuite,

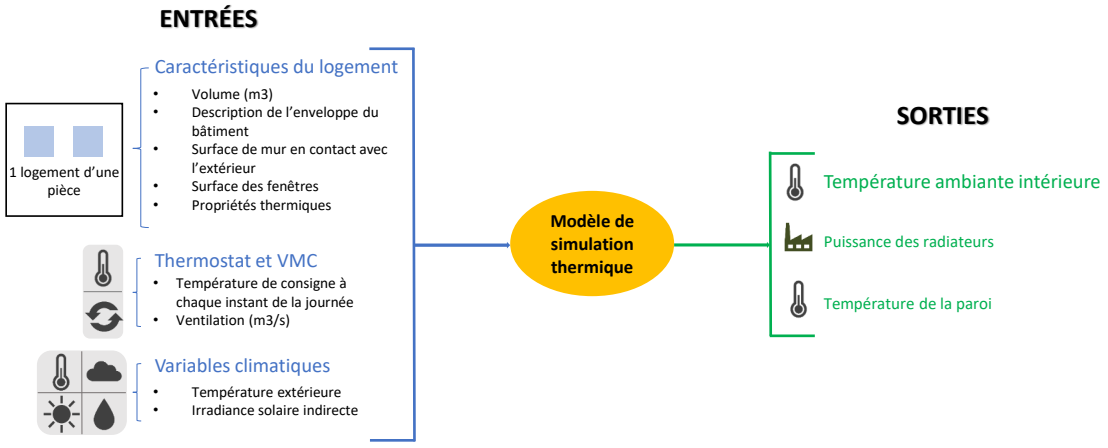

FIGURE 5.1 – Schéma récapitulatif des inputs à fournir au modèle de simulation thermique et des outputs obtenus en sortie.

on définit des profils type de température de consignes. Puis, à partir des trois profils, des signaux de thermostats sont générés pour chaque appartement.

1 - DÉFINIR LES PARAMÈTRES D'ENTRÉE DU MODÈLE

Le schéma 5.1 présente les données d'entrée nécessaire pour simuler des données de température intérieure pour un logement.

La description de l'enveloppe du bâtiment est composée de l'épaisseur des différentes couches qui forme les murs extérieurs. Les propriétés thermiques font référence à la capacité thermique volumique ainsi qu'à la conductivité thermique de chacune de ces couches. L'isolation est caractérisée par l'épaisseur des couches qui la compose, mais également par la capacité de ces couches à conserver l'énergie et transmettre la chaleur.

Les éléments suivants ont été définis pour la simulation :

Dimensions de l'appartement 40m^2 et 2,5m de hauteur sous plafond. Le volume total est de 100m^3.

Surface du mur en contact avec l'extérieur La surface de l'appartement donnant sur l'extérieur est composée de 4 m^2 de fenêtre et 16m^2 de mur.

L'isolation Le mur est du type "monomûr", composé d'une couche d'enduit intérieur, un mur porteur et une couche d'enduit extérieur, avec une épaisseur totale de 0.46m. Il s'agit d'une configuration correspondant à un logement plutôt mal isolé, mais réaliste.

La ventilation La ventilation est en marche constamment pour opérer le renouvellement de l'air. On considère ici que la ventilation mécanique contrôlée (VMC) renouvelle 6% du volume de l'appartement toutes les heures.

Capacité et conductivité thermique La capacité thermique volumique est la capacité d'un matériau à stocker l'énergie par rapport à son volume. La conductivité thermique

correspond à la capacité d'un matériau à conduire et à transmettre la chaleur. Ainsi, plus le coefficient est faible, plus le matériau est isolant. Pour finir, la résistance thermique est l'indicateur couramment utilisé pour caractériser l'isolation d'un logement. Il est obtenu en divisant l'épaisseur de la paroi par la conductivité thermique. La résistance correspond à la capacité d'un matériau à ne pas conduire la chaleur. Plus la résistance de la paroi est importante, plus cette dernière est isolante.

Ces caractéristiques sont décrites dans le tableau 5.1 pour chacune des trois couches composant la paroi de l'appartement :

TABLE 5.1 – Descriptif de l'isolation choisie pour la simulation thermique des données.

	Couche 1 Enduit intérieur	Couche 2 Mur Porteur	Couche 3 Enduit extérieur
Épaisseur (m)	0.03	0.4	0.03
Conductivité thermique (W/mK)	0.8	0.4	0.8
Résistance thermique (m^2K/W)	0.0375	1	0.0167
Capacité thermique volumique (KJ/m^3K)	1500	1200	2000

Ainsi, on obtient une résistance thermique de la paroi de 1.0542 m^2K/W, ce qui correspond à une paroi faiblement isolante. À titre de comparaison, la règlementation thermique de 2012 préconise une résistance thermique de 2.9 m^2K/W pour les murs extérieurs des logements en région parisienne.

Capacité thermique de la zone La capacité thermique volumique correspond à la capacité d'un matériau à capturer de la chaleur, rapporté à son volume. Celle de l'air vaut 1,211 KJ/m^3K. Cependant, il faut prendre en compte la présence de meubles au sein d'un appartement. Par conséquent, la capacité thermique globale de la zone avec mobilier vaut 2,6 KJ/m^3K.

Conditions météorologiques La température extérieure et l'irradiance solaire indirecte sont utilisées pour la simulation. Il s'agit de données de la zone de Trappes, correspondant au climat parisien. Les données utilisées sont des données du mois de janvier 2018. Ces données sont présentées dans la figure 5.2.

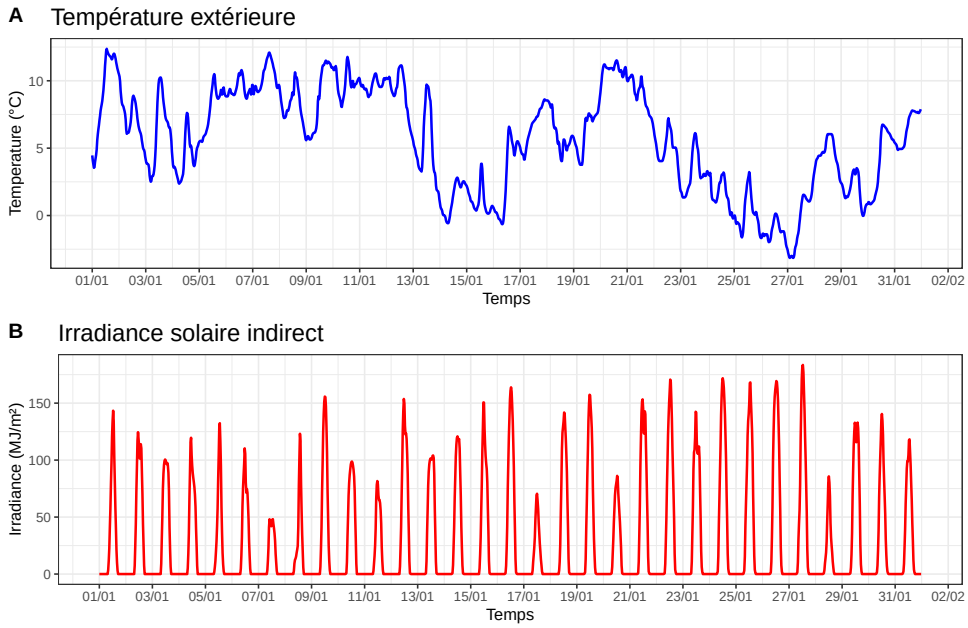

FIGURE 5.2 – Température extérieure et irradiance solaire de la zone de Trappes au cours du mois de janvier 2018.

Les éléments présentés ci-dessus seront utilisés. En plus de ces données, supposées identiques pour l'ensemble des logements, il faut générer les températures de consigne propres à chaque logement.

2- GÉNÉRATION DES TEMPÉRATURES DE CONSIGNE EN DEUX ÉTAPES

2a - Définir des profils types de chauffage

L'objectif est la construction d'un jeu de données de températures intérieures variées issues d'un ensemble de 15 appartements afin d'appliquer les modèles de classification présentés précédemment. Pour cela, il faut définir des séquences de température de consigne pour chacun de ces appartements. Tout d'abord, on se place dans un cadre dans lequel il existe 3 classes parmi les appartements. Ces clusters font référence à des comportements liés aux habitudes de chauffage et de présence. Le tableau 5.2 présente, de manière succincte, ces trois profils type.

À partir de ces comportements types, présentés dans le tableau 5.2, trois signaux de thermostat type sont construits pour une durée d'un mois et sont présentés dans la figure 5.3.

2b -Générer les séquences de température de consigne pour les 15 appartements

On part du principe que la séquence de température de consigne d'un appartement résulte d'une combinaison de trois signaux de thermostats issus des signaux type tout en conservant une classe majoritaire.

Tout d'abord, afin de générer un signal de température du consigne issu d'un signal type, on fait varier légèrement les températures de confort et les heures de changements de température demandée pour créer de la variabilité tout en gardant un profil commun. Le schéma 5.4 présente la démarche utilisée pour générer un thermostat issu du premier profil

TABLE 5.2 – Descriptif des profils types utilisés pour générer les thermostats type pour la simulation thermique de données.

Classe 1	Semaine	Température de confort demandée le soir et le matin. Pas de consigne la nuit. Horaires « de bureau »
	Week-end	Température de confort demandée en fin d'après-midi et le matin. Absence l'après midi. Pas de consigne la nuit.
Classe 2	Semaine	Température de confort demandée du début d'après-midi au soir. Pas de consigne la nuit.
	Week-end	Température de confort demandée au cours de la journée. Pas de consigne la nuit.
Classe 3	Semaine	Température demandée tout au long de la journée. La température de confort du soir peut différer de celle de la journée. Pas de consigne la nuit
	Week-end	Température demandée tout au long de la journée. Pas de consigne la nuit

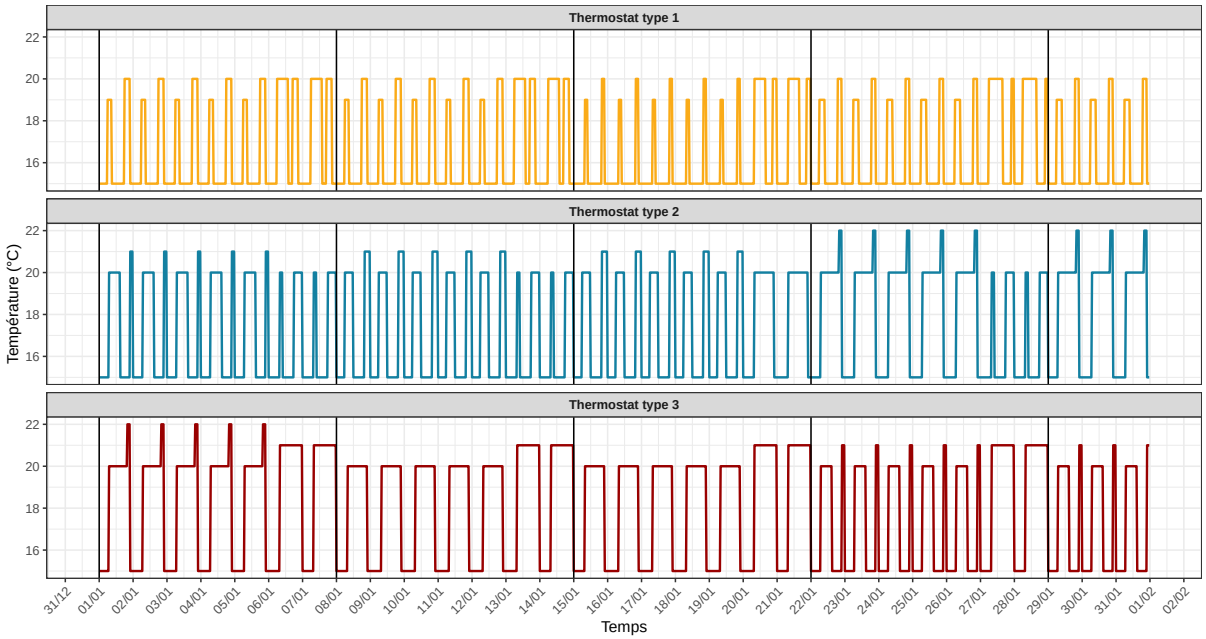

FIGURE 5.3 – Signaux de thermostats type utilisés pour générer les séquences de températures de consignes de chaque appartement pour la simulation de données thermiques. La période est de 1 mois, au pas de temps 30 minutes.

type pour une semaine.

Le principe est de tirer aléatoirement les températures de confort et les heures de changement selon des lois normales centrées sur les valeurs du thermostat type et de variances allant de 0.5 en semaine à 0.8 ou 1 pendant le week-end. Il en résulte, pour chaque appartement, trois profils issus des trois signaux de thermostat type qui sont ensuite sommés selon des poids définis en amont.

Pour simuler ces poids, on désigne d'abord une classe d'appartenance pour chaque appartement selon les effectifs choisis (6 appartements dans la première classe, 5 dans la seconde et 4 dans le troisième). Il en résulte un tableau de 3 colonnes contenant des 1 dans

les colonnes correspondantes aux classes d'appartenance (des zéros ailleurs). Ensuite, pour chaque observation, on génère des poids en bruitant, à l'aide de valeurs aléatoires comprises entre 0 et 0.5, les lignes de tableau puis en normalisant afin que la somme des poids soit égale à l'unité.

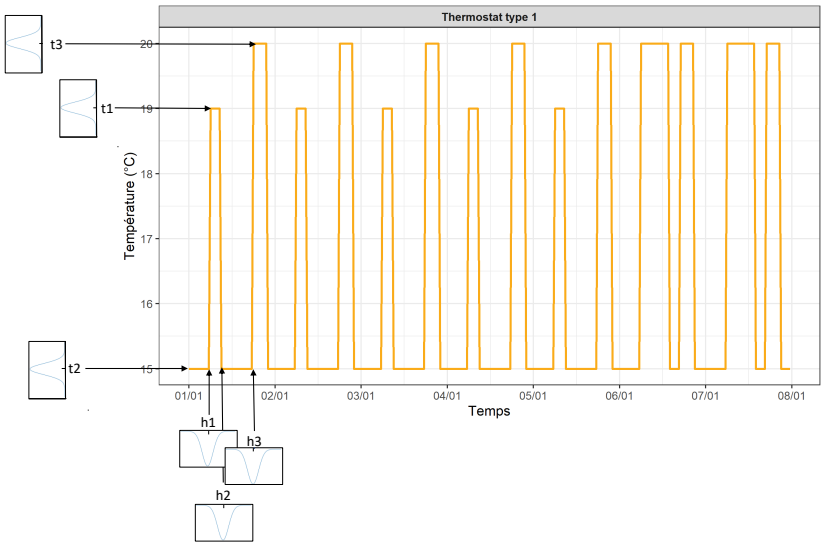

FIGURE 5.4 – Schéma descriptif de la démarche utilisée pour générer des séquences de température de consignes à partir d'un signal de thermostat type. Pour chaque heure de changement et température de confort, une nouvelle valeur est tirée aléatoirement selon une loi normale.

La figure 5.5 représente les 15 thermostats générés selon les classes d'appartenances. Ces données constituent les inputs pour le modèle de simulation thermique.

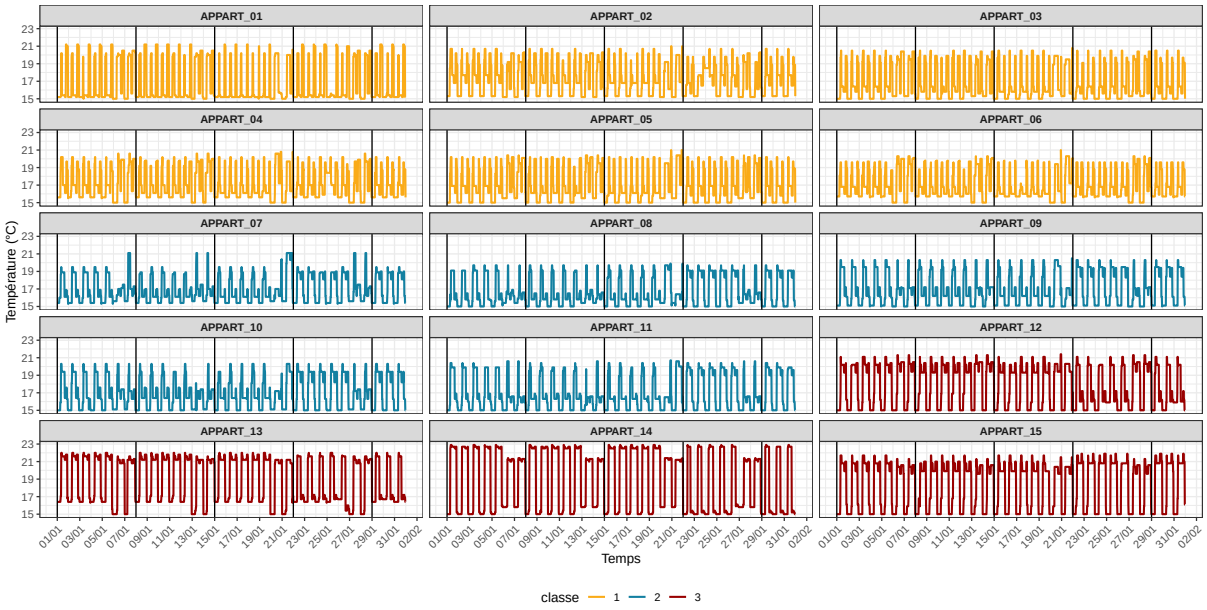

FIGURE 5.5 – Les 15 thermostats générés et utilisés comme input du modèle de simulation thermique de données de température intérieure.

3- SIMULER LES DONNÉES DE TEMPÉRATURE INTÉRIEURE À L'AIDE DU MODÈLE DE SIMULATION THERMIQUE

Finalement, à partir des caractéristiques des logements, des données météorologiques et des thermostats précédemment présentés, les données de température intérieure pour les 15 appartements sont simulées via le modèle de simulation thermique monozone pour une durée de 1 mois. Les données sont fournies au pas de temps de 5 minutes. Le graphique 5.6 représente les simulations des températures agrégées au pas de temps 30 minutes.

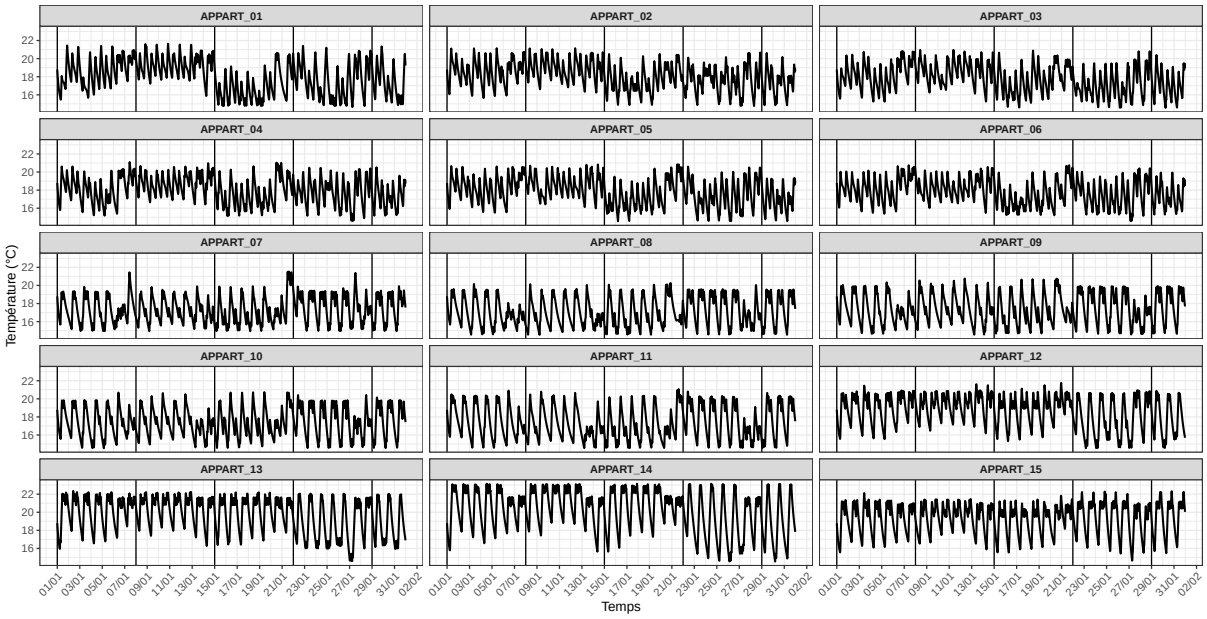

FIGURE 5.6 – Température intérieure obtenue pour 15 logements à partir du modèle de simulation thermique. Les données sont disponibles pour une période de 1 mois, au pas de temps de 30 minutes.

Par la suite, on utilise ces données pour appliquer le modèle de classification dynamique à effet exogène commun, afin d'extraire des comportements liés aux habitudes de chauffage et d'occupation.

5.2.2 Classification des données thermiques et modélisation dynamique des profils de classes pendant une semaine

Les données d'entrée, le choix des facteurs exogènes et du nombre de classes

Les données disponibles sont des températures intérieures de 15 appartements ainsi que les mesures de température extérieure et d'irradiance solaire indirecte. La figure 5.7 représente les données de température intérieure simulées pour une semaine.

Afin d'estimer le modèle de classification à effet exogène commun, il faut définir les facteurs à intégrer et le nombre de classes. Pour sélectionner ces deux éléments, on utilise le critère BIC, qu'on cherche à minimiser. Le critère BIC est largement utilisé pour la sélection de modèles. Dans la mesure où la log-vraisemblance n'est pas disponible pour ce modèle, ce

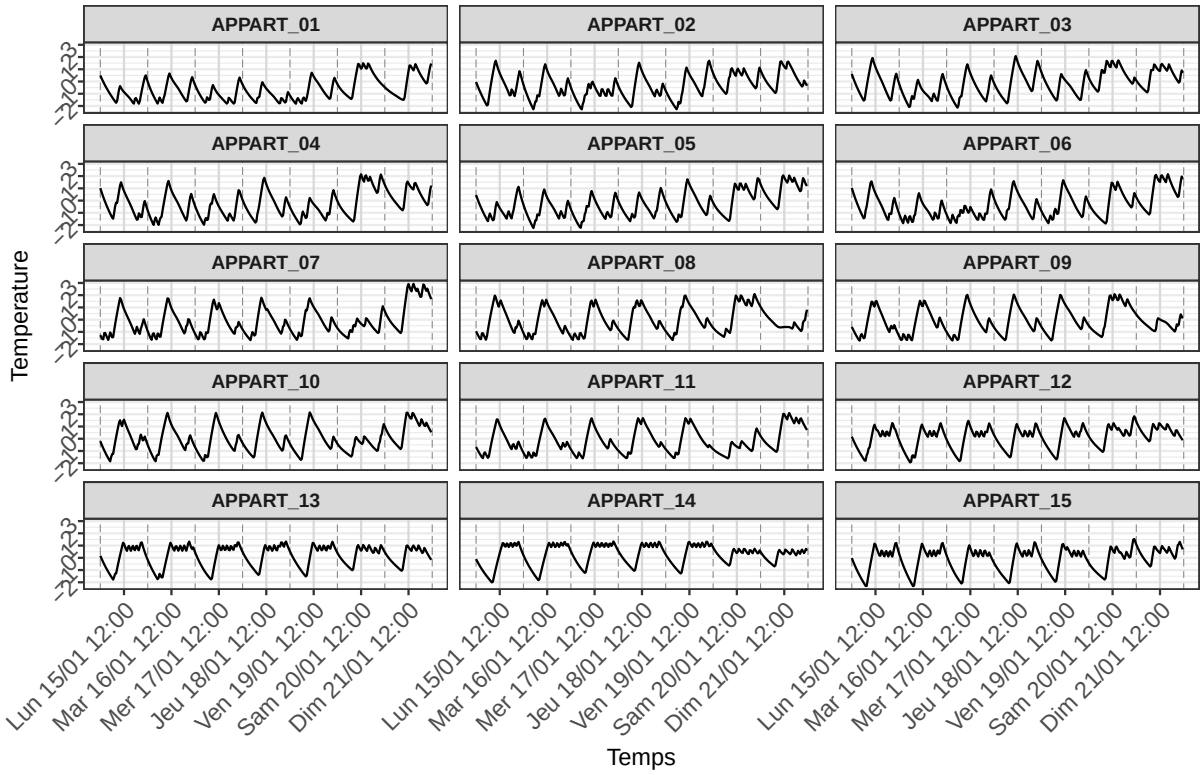

FIGURE 5.7 – Une semaine de données de température intérieure simulées, centrées et réduites pour 15 appartements à partir du modèle de simulation thermique au pas de temps 30 minutes.

critère est défini à l'aide de la borne inférieure de la vraisemblance 4.11 pour un modèle noté c tel que :

$$BIC(c) = -2F(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\tau}}, \boldsymbol{\Theta}) + N_c \log(nT), \tag{5.1}$$

où n est le nombre d'observations, T la longueur de la séquence observée et N_c le nombre de paramètres libres du modèle c.

Le modèle permet d'estimer un effet régressif relatif à des facteurs exogènes communs à l'ensemble des classes. Les données météorologiques, comme la température extérieure et l'irradiance solaire, sont d'emblée considérées comme des candidats dans la mesure où elles sont utilisées lors de la simulation. L'une des motivations applicatives lors de la construction des modèles est de séparer l'effet du contexte météorologique afin d'estimer des profils de classes ne dépendant plus de ces facteurs. De plus, l'heure de la journée est prise en compte comme un facteur exogène et est introduite sous forme de variable périodique afin de capter l'aspect périodique des comportements. La sélection de facteurs consiste à définir combien de variables horaires sont utilisées pour modéliser l'heure de la journée. Avec J désignant le nombre de variables périodiques horaires et considérant qu'on dispose de données au pas de temps 30 minutes, ces dernières sont construites telles que :

$$\forall\, t \in \{1,2,\ldots,T\}, \forall\, j \in \{1,...,J\} \begin{cases} \mathrm{x}_{jt} = \sin(\frac{2\pi jt}{48}) \\ \mathrm{y}_{jt} = \cos(\frac{2\pi jt}{48}) \end{cases} \tag{5.2}$$

À partir des variables horaires ainsi que des données de température extérieure et d'irradiance solaire, cinq ensembles de facteurs exogènes sont construits. L'étape de sélection consiste à choisir, parmi ces 5 possibilités, quels ensembles de facteurs permettent de minimiser le critère du BIC.

Les 5 ensembles de facteurs sont les suivants :

Facteurs 1 : Température extérieure, Irradiance solaire et 6 variables horaires ($J = 3$).

Facteurs 2 : Température extérieure, Irradiance solaire et 4 variables horaires ($J = 2$).

Facteurs 3 : Température extérieure, Irradiance solaire et 2 variables horaires ($J = 1$).

Facteurs 4 : Température extérieure, Irradiance solaire.

Facteurs 5 : 4 variables horaires ($J = 2$).

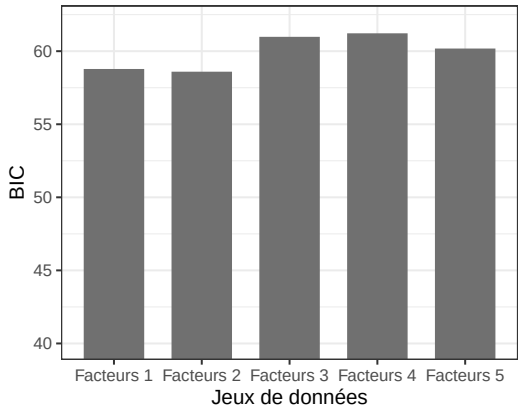

FIGURE 5.8 – BIC obtenu pour l'estimation du modèle avec différents ensembles de facteurs exogènes considérés et $K = 3$ classes construites.

La figure 5.8 représente les valeurs de BIC obtenues à l'issue de l'estimation du modèle avec les 5 ensembles de facteurs et $K = 3$ classes. Finalement, pour l'application, on utilisera l'ensemble n°2, à savoir la température extérieure, l'irradiance solaire indirecte et quatre variables horaires. La figure 5.9 représente ces facteurs sur la période d'intérêt. Le nombre de classes est fixé à $K = 3$ car l'information est disponible dans la mesure où 3 profils type ont été construits pour la simulation des données.

Classification et interprétation des résultats

Le modèle de classification dynamique avec estimation d'un effet exogène commun est appliqué sur les données de température d'une semaine, au pas de temps 30 minutes pour construire 3 classes. De plus, les facteurs exogènes sont composés de la température extérieure et de l'irradiance solaire qui ont été centrées et réduites, et de 4 variables périodiques modélisant l'heure de la journée.

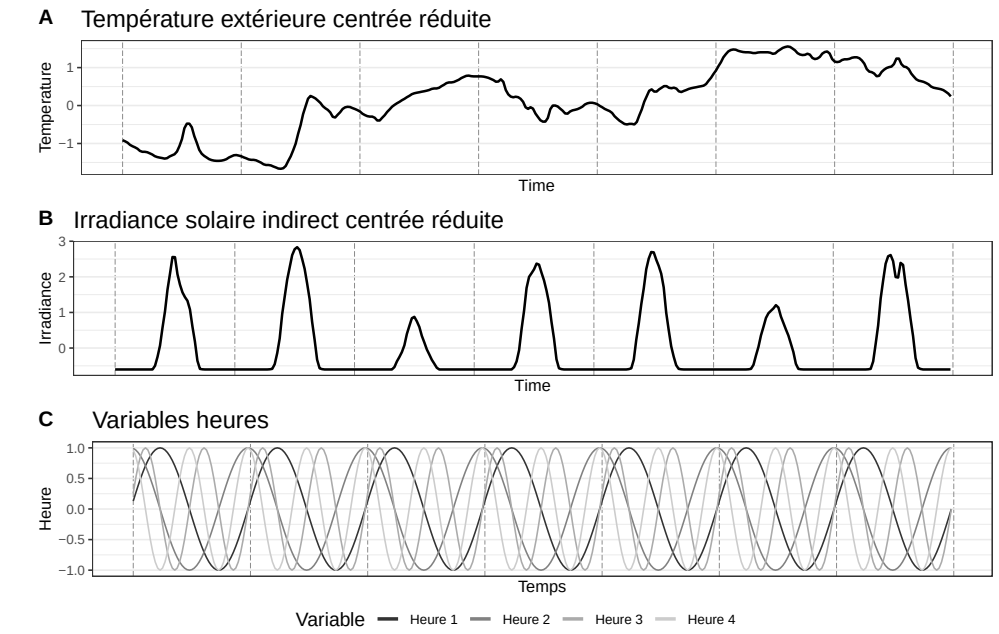

FIGURE 5.9 – Facteurs exogènes considérés pour la construction des clusters et l'estimation des profils de classes pendant une semaine, au pas de temps 30 minutes. (A) représente la température extérieure centrée réduite. (B) représente l'irradiance solaire indirecte, centrée et réduite. (C) représente quatre variables périodiques représentant l'heure de la journée.

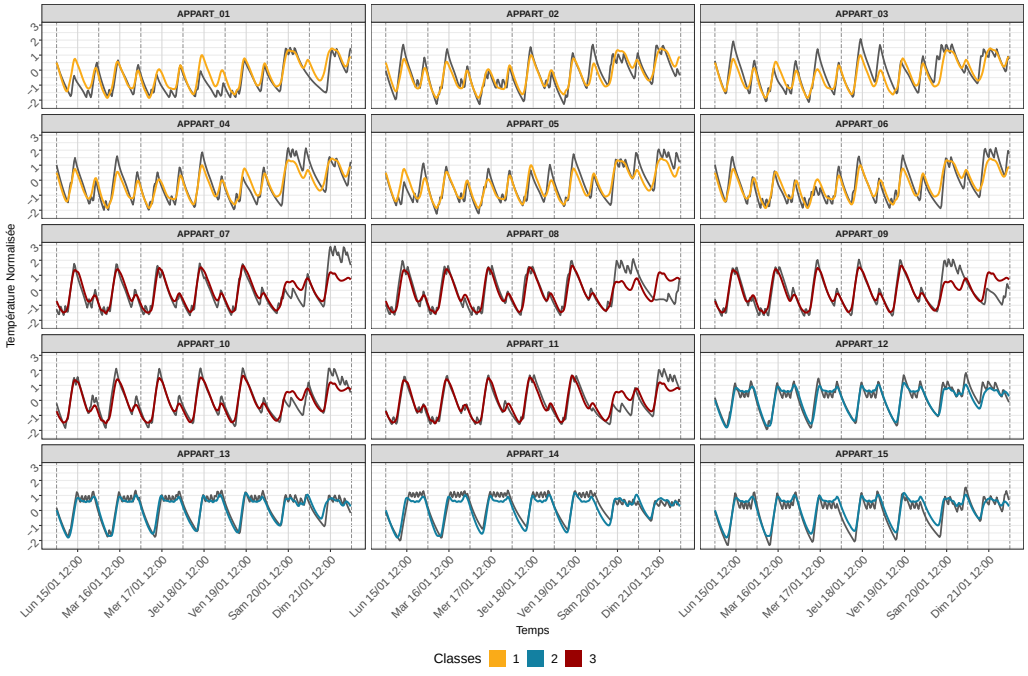

FIGURE 5.10 – Données de température centrées réduites simulées à partir du modèle thermique et estimations fournies par le modèle de classification selon la classe d'appartenance estimée.

Les données estimées, présentées dans la figure 5.10 sont obtenues comme suit :

$$\forall i = 1,\dots,n \text{ et } t = 1,\dots,T \quad \hat{x}_{it} = \mathbf{u}_t'\hat{\mathbf{a}} + \sum_{k=1}^{K} \hat{\tau}_{ik}\hat{b}_{kt}. \tag{5.3}$$

Tout d'abord, le premier élément à relever des résultats de la figure 5.10 est que les classes estimées correspondent à celles initialement simulées. De plus, on peut noter une plutôt bonne adéquation des données estimées aux séquences simulées. La variable $\hat{z}_i$ dénote la classe à laquelle appartient l'observation i qui est définie par la règle du maximum *a posteriori* (MAP).

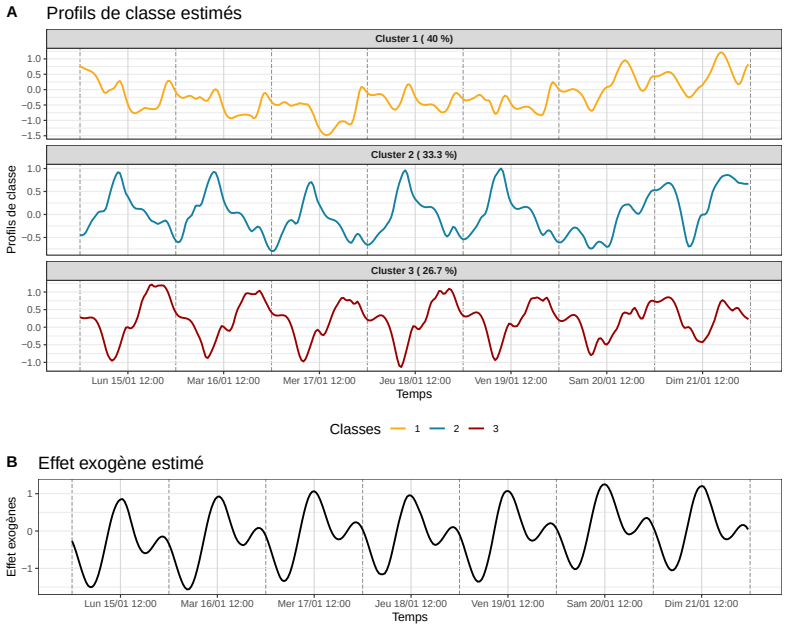

FIGURE 5.11 – Profils de classes et effet exogènes estimés via le modèle de classification à partir des données issues de la simulation thermique.

Ensuite, la figure 5.11 représente les centres de classes estimés (5.11A) par le modèle et l'effet régressif commun (5.11B). On peut observer que l'effet régressif est fortement lié aux variables périodiques horaires et indique deux pics journaliers considérés comme communs à l'ensemble des appartements. On observe également un niveau légèrement plus élevé le week-end.

Ensuite, à partir de la figure 5.11A, on peut noter que le profil n°1 ne présente pas de pics apparents en semaine. Cependant, au cours du week-end, on observe un pic en début d'après-midi. Concernant, le second profil, on observe des pics hauts en fin de matinée et des pics bas au milieu de la nuit ainsi qu'un étalement le samedi. Pour finir, le troisième profil présente des hauts et longs pics en milieu d'après-midi. Les périodes de température basses sont courtes et le pic bas arrive en début de matinée avant de remonter tout au long de la journée.

Afin de comprendre et d'interpréter ces profils, ces derniers sont superposés aux thermostats types présentés dans la figure 5.3. Ainsi, la figure 5.12 représente les trois profils de classes superposés aux thermostats types correspondants. Les classes ont été réorganisées afin de correspondre aux classes simulées.

Sur la figure 5.12 on observe un lien fort entre les thermostats types et les centres de

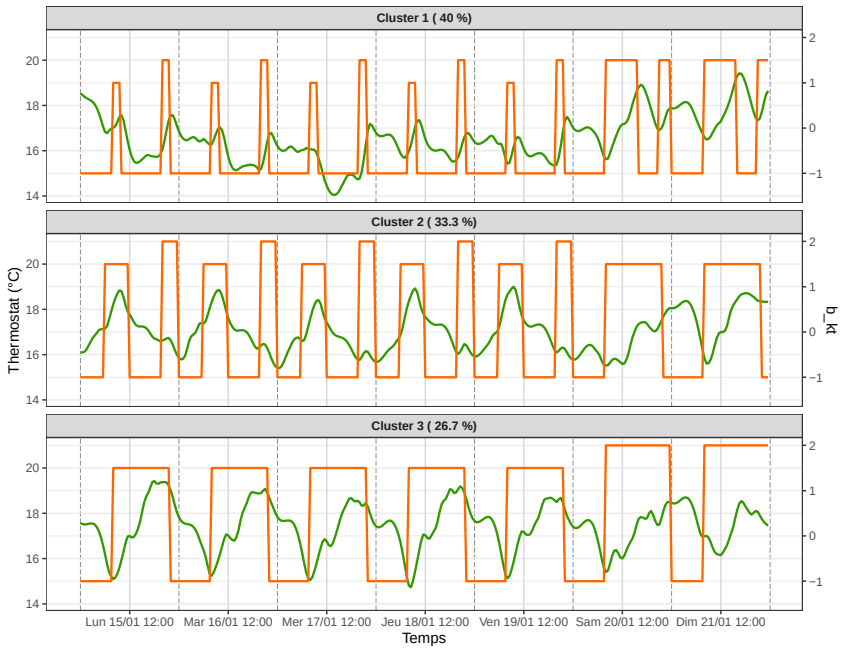

FIGURE 5.12 – Superposition des profils de classes estimés et des thermostats types utilisés pour la simulation des données de température.

classes estimés. Sur le thermostat n°1, on observe deux pics de température de consigne le matin et le soir. Cependant, la durée de demande est finalement très courte. Par conséquent, on ne retrouve pas ces pics sur les centres de classes dans la mesure où la température n'a pas le temps d'atteindre la température demandée avant que, finalement, une nouvelle consigne plus faible ne soit demandée. Ce phénomène est dû au temps nécessaire au chauffage de permette à la température de la pièce d'augmenter. On obtient donc des profils de classes relativement constants au cours de la semaine. Le week-end par contre, les périodes de demande de température de confort sont plus longues, donc ces périodes se manifestent par des pics au cours de l'après-midi dans les profils de classes.

Pour le second thermostat, les périodes de demande de températures sont plus longues au cours de la matinée. On observe cet effet sur les centres de classes qui forment des pics hauts au cours de la matinée. Cependant, le second pic du soir, plus court, ne se retrouve pas forcément sur les profils des classes, pour les mêmes raisons évoquées pour le premier cluster. De plus, on peut noter que les pics bas des profils de classes coïncident avec l'instant où la consigne de température change pour passer d'une température d'absence à la température de confort.

Pour finir, le troisième thermostat correspond à des personnes qui ont une température de consigne élevée tout au long de la journée sauf la nuit. On observe donc de longs pics sur le centre de classe correspondant à la température intérieure qui augmente tout au long de la journée pour atteindre la consigne, puis qui diminue tout au long de la nuit. Ainsi, le pic bas coïncide avec le moment où la température de consigne est demandée.

Ces observations permettent de confirmer que les profils de classes estimées reflètent

les profils types initialement simulés. Cela permet aussi de confirmer que le modèle de classification permet de construire des classes homogènes de logement à partir de données de températures et d'estimer les dynamiques comportementales relatives aux habitudes de chauffage.

5.2.3 Séparation des effets endogènes aux comportements, des effets régressifs exogènes

Le modèle de classification avec estimation des centres de classes dynamiques et effet régressif commun est motivé par la volonté de regrouper des observations sous forme de clusters tout en estimant des effets régressifs relatifs à des facteurs connus et en modélisant la partie restante et inexpliquée sous forme de processus autorégressifs. Par conséquent, l'un des aspects intéressants de ce modèle réside dans la séparation des effets dits endogènes et exogènes. Par exemple, dans un problème de régression, l'objectif est d'estimer l'effet d'un ensemble de facteurs exogènes en supposant que le reste, la partie non expliquée, est un bruit gaussien. Finalement, on pourrait imaginer que cette partie inexpliquée n'est pas uniquement un bruit, mais le résultat de facteurs non observés que l'on cherche également à modéliser.

Le modèle de classification avec estimation d'un effet régressif est donc appliqué à une séquence d'une semaine pour un appartement simulé. La figure 5.13 représente les données d'entrée utilisées pour l'estimation (5.13(A)), ainsi que l'effet exogène estimé à partir de la température extérieure normalisée, l'irradiance solaire normalisée et 4 variables périodiques représentant l'heure de la journée (5.13(B)) et le profil de classe estimé (5.13(C)).

L'effet des facteurs exogènes capte les pics de la température en fin de matinée pour le plus petit pic et le soir pour le plus important. La partie non expliquée par ces facteurs est estimée par le profil de classe. On peut noter que l'effet périodique et régulier de l'occurrence des pics est capté par l'effet exogène grâce aux variables horaires. Ainsi, le profil de classe (C) capte les variations qui ne sont pas périodiques et régulières. Par exemple, on peut noter un niveau de température intérieure légèrement plus élevé le mardi et le début de la journée du mercredi sur le graphique (A). On retrouve ce niveau général plus important sur le profil de classe à la même période. L'effet exogène capte les deux pics journaliers d'intensité voisine observée au cours de la matinée et le soir. Durant le week-end, les périodes de demande de chauffage sont plus longues, le profil de classe capte les changements qui ne sont pas pris en compte par l'effet régressif. Les pics observés sur le profil de classe semblent s'allonger au cours du week-end (deux derniers jours).

La figure 5.14 représente les 15 profils de classes et les 15 effets exogènes estimés suite à l'application du modèle sur chaque séquence d'appartement. On peut noter que les effets régressifs sont très semblables au sein de chaque classe. Cela peut confirmer les premières observations selon lesquelles l'effet régressif capterait l'aspect périodique des comportements qui, dans ce cas de la simulation, sont communs aux appartements d'une même classe. En

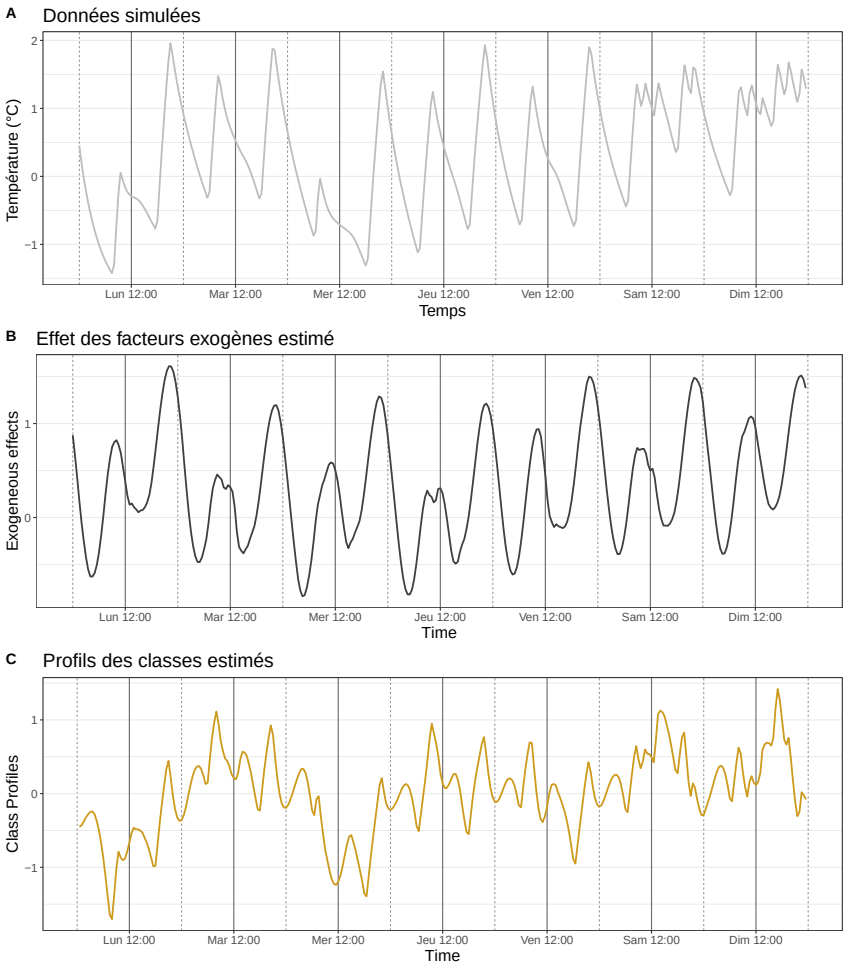

FIGURE 5.13 – Données d'entrée, profil de classe et effet exogène estimé via le modèle appliqué à une semaine de données de l'appartement n°1. (A) représente les températures normalisées d'entrée et l'estimation, (B) représente l'effet des facteurs de température, d'irradiance et d'heure, et (C) représente le profil de classe estimé.

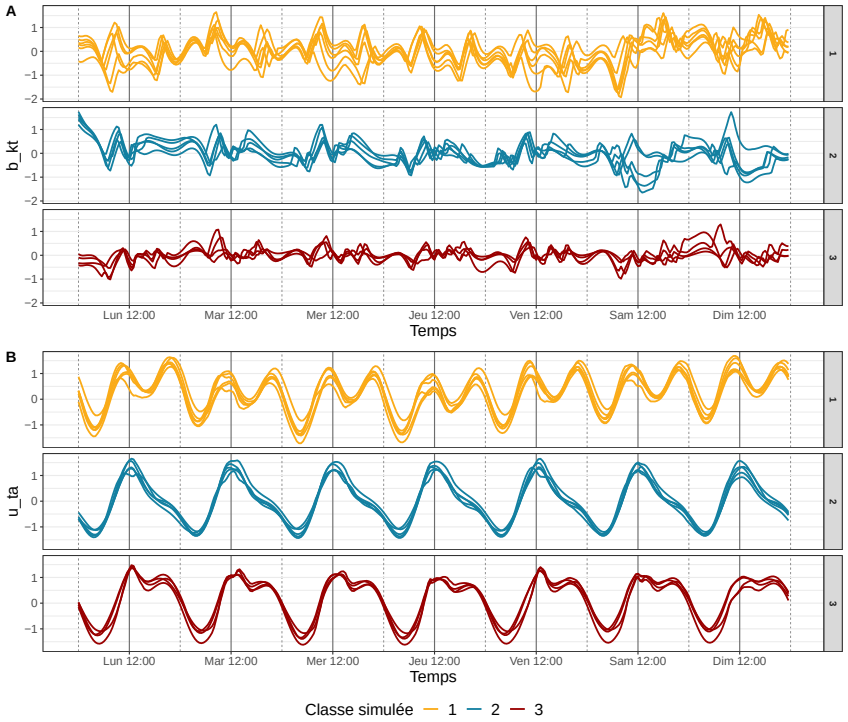

FIGURE 5.14 – Profil de classe et effet régressif estimé pour chaque appartement individuellement selon la classe d'appartenance utilisée pour la simulation

effet, les thermostats utilisés pour la simulation sont issus de thermostats types faisant référence à des périodes de demande de température, de confort en fonction d'horaires de présence et d'absence. Les profils de classes estimés, bien que semblables au sein de chaque classe, sont légèrement plus variés et font référence à la variabilité propre à chaque appartement. De plus, les changements de comportements le week-end pour l'ensemble des appartements s'observent sur les profils de classe, car l'aspect périodique des variables heures ne permet pas de prendre en compte ces changements de comportements. En d'autres termes, les changements de comportements au cours du week-end ne sont pas captés par l'effet exogène qui est lié, en partie, à l'heure de la journée, car le comportement n'est pas le même qu'au cours de la semaine. Les profils de classes permettent donc de capter des variations de comportements au cours de la semaine. Pour finir, on peut observer que les effets régressifs estimés pour chaque appartement sont assez différents d'une classe à l'autre. En effet, l'occurrence et l'intensité des pics ne sont pas les mêmes. Ces observations constituent une motivation supplémentaire à l'utilisation du modèle avec estimation d'effets régressifs propres à chaque cluster.

Ces premiers résultats obtenus sur des données simulées à partir d'un modèle de simulation thermique permettent de confirmer le lien entre les profils de classes estimé et les comportements de chauffage des habitants. Par la suite, on applique le modèle à des données réelles pour identifier des classes de maisons dont les habitants ont des comportements similaires.

5.3 Application à des données thermiques réelles issues d'un ensemble de maisons individuelles

Le jeu de données [Firth et al., 2017], qu'on appellera REFIT, utilisé dans cette application est constitué de 20 maisons individuelles situées à Loughborough, au Royaume-Uni. Ce jeu de données contient, entre autres, la température intérieure des pièces à vivre des maisons, ainsi qu'un ensemble de caractéristiques concernant les habitants des logements, et enfin des données météorologiques telles que la température extérieure ou l'irradiation solaire. Des données de détection de mouvement sont également disponibles pour dix-neuf des maisons pendant des périodes relativement courtes. Plus de détails sont fournis dans le chapitre 2.

Les objectifs de l'application du modèle de classification sur les données réelles sont les suivants :

- Construire des clusters regroupant des maisons ayant des dynamiques de températures similaires sur une période donnée.
- Estimer l'effet régressif des facteurs météorologiques et de l'heure de la journée sur les températures au sein des maisons.
- Estimer les centres de classes comme des processus autorégressifs caractérisant les comportements endogènes des habitants.
- Interpréter et caractériser les classes et les profils de classes estimés à l'aide de variables de détection de mouvement, de consommation d'électricité ou de variables socio-démographiques.

Pour mener à bien ces objectifs, il faut, dans un premier temps, sélectionner la période d'intérêt, les facteurs exogènes à intégrer et le nombre de classes à construire à partir du critère du BIC. Ensuite, une fois le modèle estimé, les classes et les profils de classes sont analysés et un lien avec la présence d'habitants au sein du logement peut-être établi. Nous appliquons le modèle de classification avec effet régressif commun puis, dans un second temps, le modèle avec effets régressif propres à chaque classe.

5.3.1 Application du modèle de classification avec centres dynamiques et effet exogène commun sur des données d'une semaine

Le choix des données, des facteurs et du nombre de classes

La sous-section suivante est consacrée à l'application du modèle sur un jeu de données d'une semaine de température ambiante au sein de 18 maisons. La période choisie se situe entre le 24 et le 30 novembre 2014, car des données de détection de mouvement sont également disponibles pour cette période. En contrepartie, les données de température pour la maison 2 et la maison 17 ne sont pas disponibles. Ces variables de détection de mouvement

fournissent des informations supplémentaires qui sont utiles pour confirmer l'interprétation des profils de classes. La figure 5.15 représente les entrées du modèle. On dispose de 18 séquences de 336 observations (T=336, n=18) de températures intérieures normalisées.

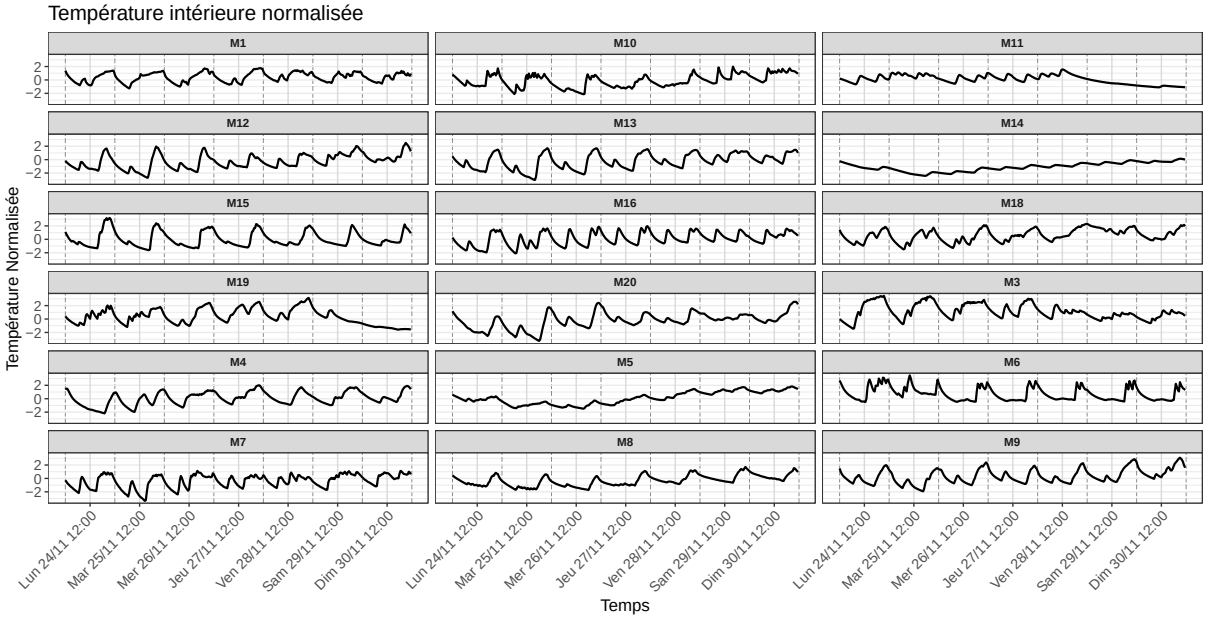

FIGURE 5.15 – Températures intérieures mesurées et normalisées au sein de 18 maisons au cours de la semaine du 24 novembre au 30 novembre 2014.

Afin d'estimer les paramètres du modèle de classification à centres de classes dynamiques et à effets exogènes communs présenté dans le chapitre 3, il faut d'une part définir les facteurs exogènes à intégrer et le nombre de classes. Le critère BIC, défini par l'équation (5.1), est utilisé pour sélectionner ces éléments.

Sélection des facteurs exogènes utilisés pour estimer l'effet régressifs communs

Le modèle permet d'estimer un effet régressif commun relatif à des facteurs exogènes. Les données météorologiques comme la température extérieure et l'irradiance sont de bons candidats. Il s'agit des deux facteurs utilisés comme entrées lors de la simulation de données thermiques, dans la première partie de ce chapitre. Comme pour l'application précédente, on considère cinq ensembles de facteurs candidats pour l'estimation d'un effet exogène commun. Les ensembles de facteurs sont présentés dans la section 5.3.1.

La figure 5.16 représente les valeurs de BIC obtenues lors de l'estimation du modèle avec les 5 ensembles de facteurs avec $K = 5$ le nombre de classes. Finalement, pour l'application, on utilisera l'ensemble n°3, à savoir la température extérieure et l'irradiance solaire normalisée et deux variables horaires. La figure 5.17 représente ces facteurs sur la période d'intérêt. Pour définir l'ensemble de facteurs, on fixe le nombre de classe à 5 car ce nombre permet de construire des classes contenant plusieurs maisons avec au maximum une seule classe vide pour les différents ensembles de facteurs.

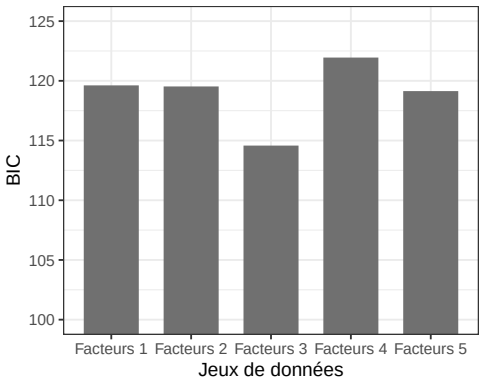

FIGURE 5.16 – Critère BIC calculé pour les 5 ensembles de facteurs exogènes.

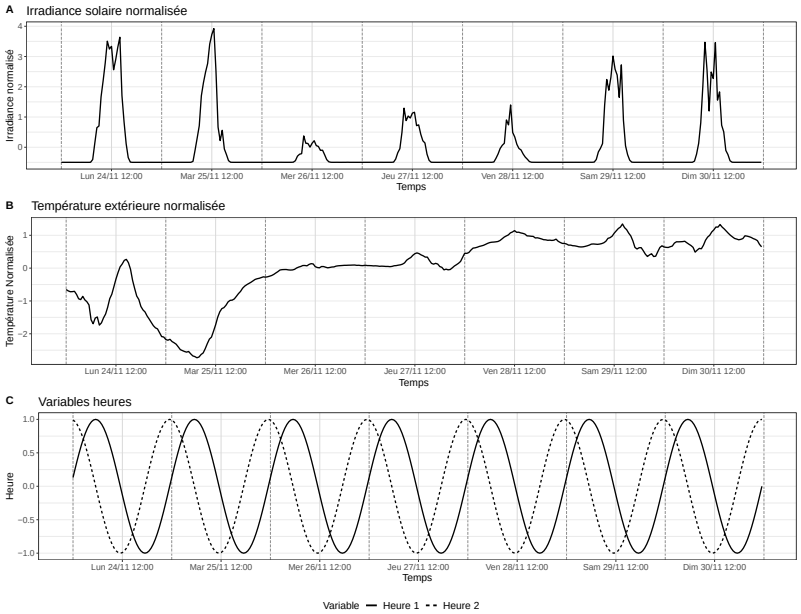

FIGURE 5.17 – La température extérieure normalisée (A), l'irradiance solaire normalisée (B) et les variables horaires périodiques (C) pendant une semaine, au pas de temps 30 minutes utilisées comme facteurs exogènes.

Le nombre de classes

Ensuite, la figure 5.18 représente les valeurs BIC obtenues pour $K = 2, ...17$ en utilisant l'ensemble de facteurs n°3 choisi à l'étape précédente.

Le critère BIC diminue jusqu'à $K = 16$. Cependant, l'objectif est de construire un petit nombre de clusters pour résumer les comportements et interpréter les clusters. De plus, pour $K > 5$, les clusters supplémentaires ne contiennent qu'une seule maison. Par conséquent, on fait le choix de construire 5 clusters.

Classification des données et interprétation des résultats

Les résultats suivants ont été obtenus à partir de l'estimation du modèle proposé sur les données (voir figure 5.15) avec $K = 5$ clusters. Le modèle estime les centres de classes ($\hat{\mathbf{b}}$),

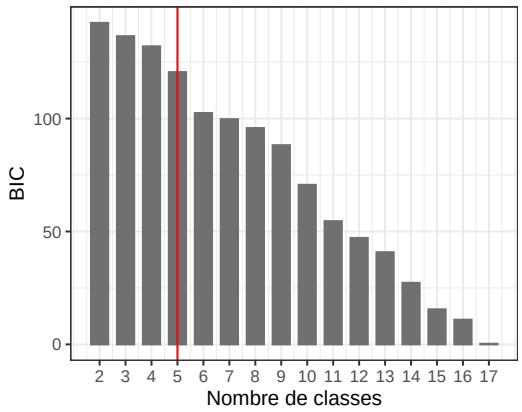

FIGURE 5.18 – Critère BIC calculé lorsque le nombre de clusters K varie entre $K = 2, ..., 17$.

l'effet exogène ($\hat{\mathbf{a}}$) et aussi, pour chaque observation et cluster, la probabilité d'appartenance $(\hat{\tau}_{ik})_{i,k}$). En utilisant ces éléments, les données peuvent être reconstruites à partir de la formule (5.3). Pour rappel, la classe d'appartenance des observations est estimée à partir de la méthode du MAP. La figure 5.19 montre les observations d'entrée et données estimées en fonction de la classe à laquelle appartient chaque appartement.

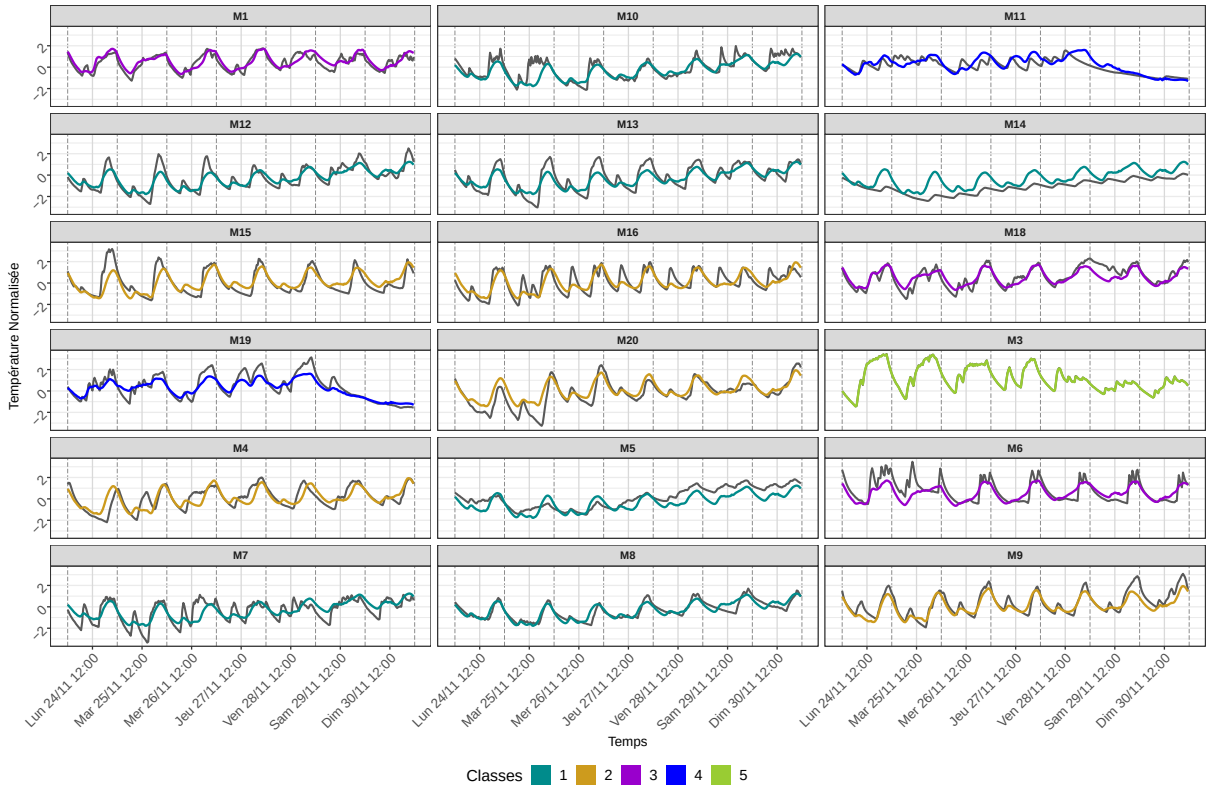

FIGURE 5.19 – Température intérieure normalisée réelle pour dix-huit maisons. En superposition, les courbes, colorées en fonction de la classe à laquelle appartiennent ces maisons, représentent les données estimées via le modèle proposé avec K=5. Les données estimées sont obtenues en utilisant l'équation (5.3).

Sur la base des résultats présentés dans la Figure 5.19 et la Figure 5.20, nous observons

tout d'abord que le cluster 5 ne comprend qu'une seule maison. De même, si l'on considère la dynamique des quatre autres clusters, on peut observer certaines différences. Ni les pics élevés ni les pics bas ne se produisent au même moment. Il existe également des différences dans le nombre et la forme de ces pics. Pour le cluster 2, par exemple, nous pouvons observer un pic élevé à la fin de chaque journée et d'autres pics plus petits à la fin des matinées pendant les jours ouvrables (du lundi au vendredi). Pour le cluster 3, les pics élevés semblent être plus longs que pour les clusters 1 et 2. Pour le cluster 4, aucune variation significative des pics ne peut être observée. De plus, la dynamique change au cours du week-end pour tous les clusters.

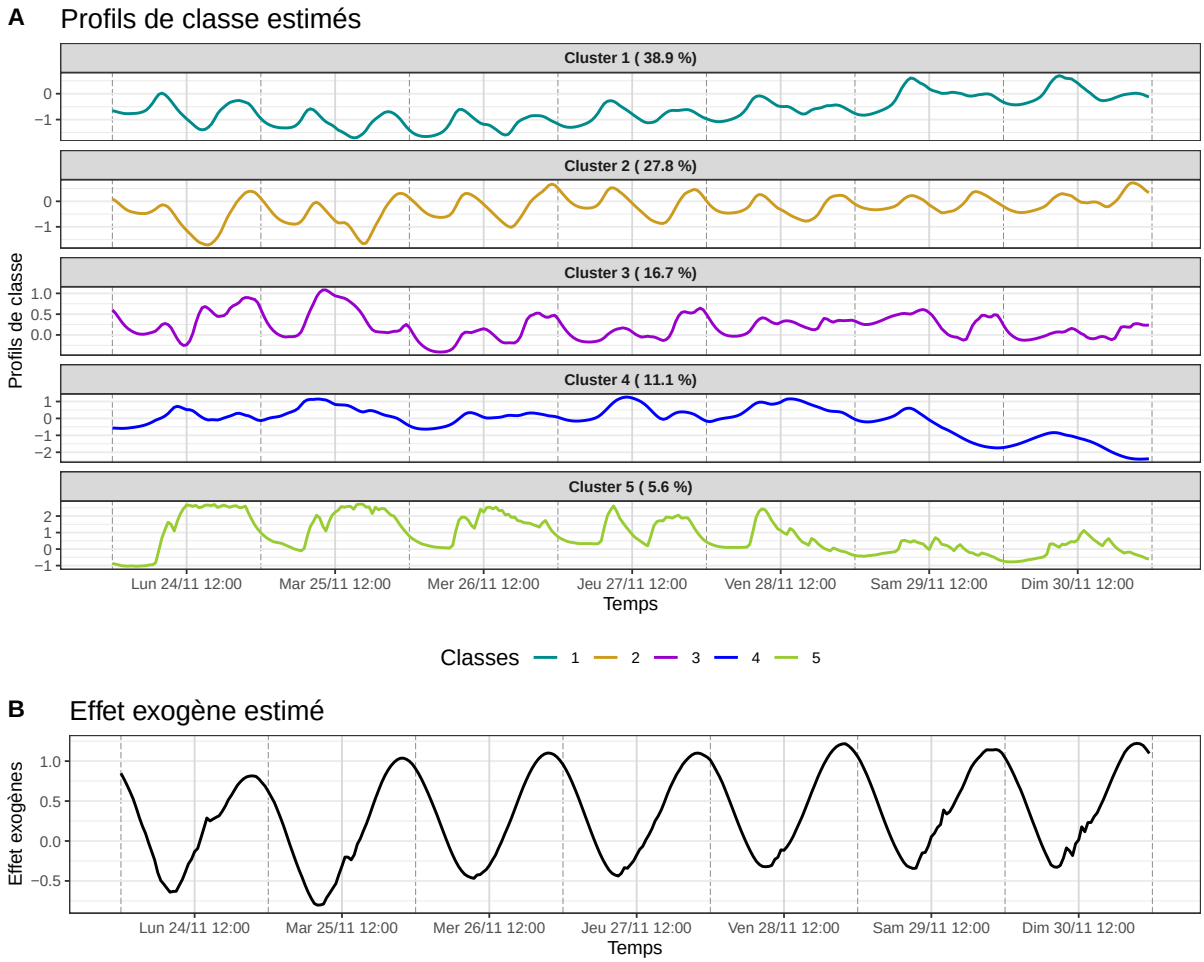

FIGURE 5.20 – Profils de classes estimés (A) et effet exogène estimé (B). Le graphique (A) représente les profils de classes estimés à l'aide du modèle proposé. Les étiquettes des clusters ont été réorganisées en fonction des proportions des classes. Le graphique (B) représente l'effet estimé des facteurs exogènes. Ces résultats ont été obtenus en utilisant les données météorologiques normalisées et les variables horaires multipliées par les paramètres de régression estimés.

Cette période a été choisie car des données de détection de mouvement sont également disponibles. En effet, la dynamique des clusters peut être liée aux comportements d'occupation. Les données de détection de mouvement ont donc été utilisées pour confirmer cette hypothèse et interpréter les profils. La figure 5.21 représente la dynamique estimée des classes

différenciée et le nombre moyen de détections dans les maisons pour chaque cluster. Comme on peut le voir, il existe, a priori, un lien entre le nombre de détections de mouvement et le profil de classe identifié. La durée du pic observé sur les profils de classes estimés peut être reliée à la période pendant laquelle le nombre de mouvements observés est élevé. De la même manière, l'heure des pics semble être liée aux heures des pics de présence avec peut-être un léger décalage dû à l'inertie de la température par exemple. Ce lien est encore plus évident pour le profil 4 qui diminue pendant le week-end. Cette baisse correspond à une présence très faible sur la même période.

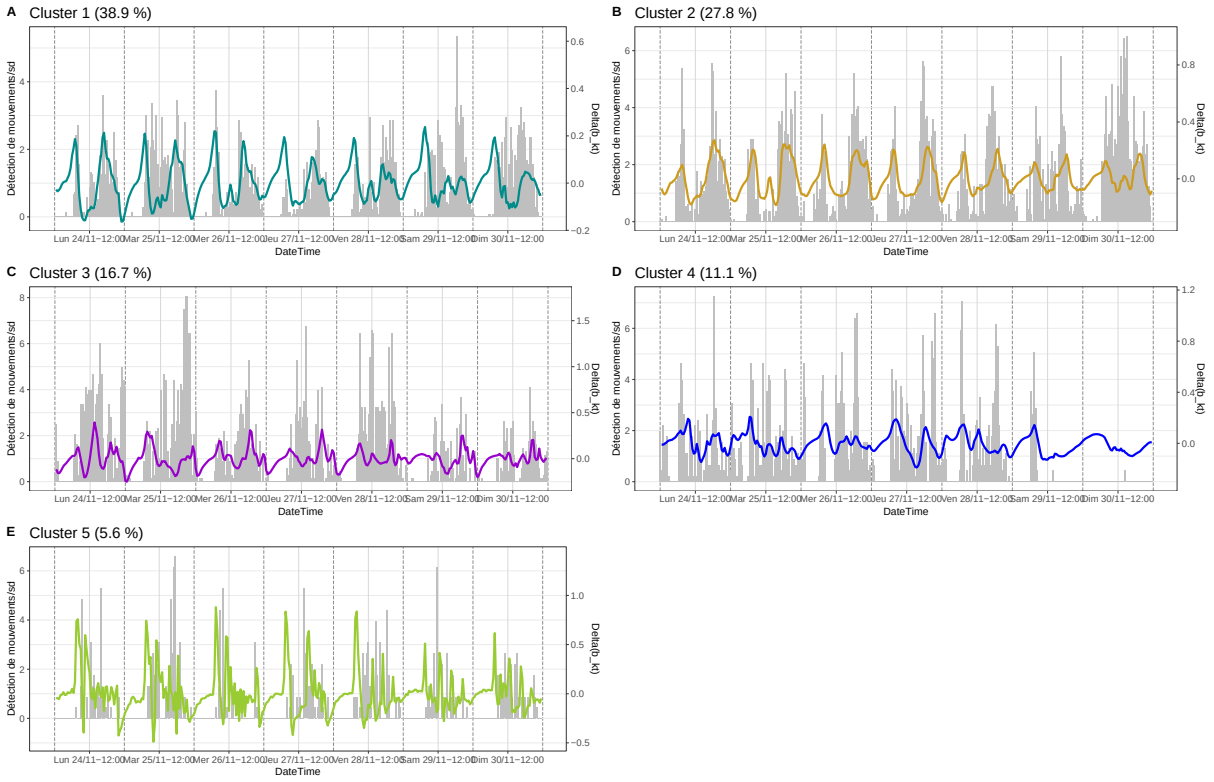

FIGURE 5.21 – Profils des classes différenciés et nombre moyen de détections de mouvement pour chaque cluster pendant la période du 24 au 30 novembre 2014. Les graphiques affichent, en gris, le nombre moyen de détections de mouvement pour chaque cluster. Afin d'ajuster les profils de classes et les données de détection de mouvement, ces dernières ont été divisées par l'erreur standard.

La figure 5.22 représente la consommation d'électricité moyenne et les profils de classes estimés et différenciés. On peut observer un lien entre ces deux variables. En effet, les variations des profils de classes semblent liées à la consommation d'électricité. Pour rappel, les maisons disposent d'un chauffage au gaz, donc la consommation d'électricité est liée à la lumière, l'utilisation d'appareils ou éventuellement l'utilisation d'un chauffage d'appoint. Ces usages peuvent être en grande partie liés à la présence des habitants. Cela renforce l'observation selon laquelle les profils de classes ont un certain lien avec le nombre de détections de mouvements.

Pour confirmer cette observation, le tableau 5.3 montre les coefficients de corrélation linéaire calculés entre les profils de classes de chaque groupe et le nombre moyen de dé-

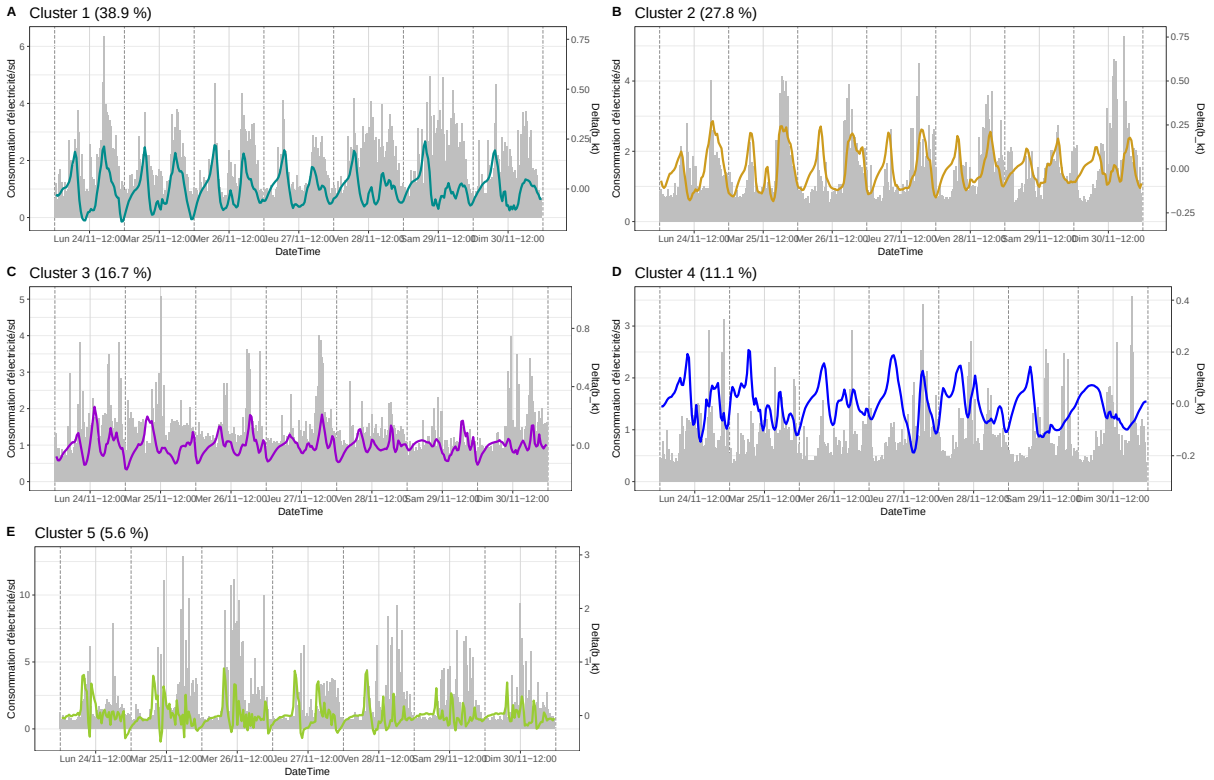

FIGURE 5.22 – Profils de classes différenciés et consommation d'électricité moyenne pour chaque cluster pendant la période du 24 ou 30 novembre 2014.

TABLE 5.3 – Coefficients de corrélation linéaire et p-values résultant du test d'indépendance. Le coefficient r est la corrélation linéaire : $r = cov(x,y)/s_x s_y$, avec s_x et s_y, respectivement, l'erreur standard calculée sur les échantillons x et y. La valeur p est le résultat d'un test statistique avec comme hypothèse nulle, la nullité de la corrélation.

$(b_{kt}, \text{motion}_{kt})$					
Cluster	**1**	**2**	**3**	**4**	**5**
r	0,298	0,013	0,125	0,416	0,299
p.value	$2{,}469e^{-08}$	0,815	$2{,}231e^{-02}$	$1{,}717e^{-15}$	$2{,}253e^{-08}$
$(\Delta(b_{kt}), \text{motion}_{kt})$					
Cluster	**1**	**2**	**3**	**4**	**5**
r	0,173	0,575	0,116	0,216	-0,060
p.value	$1{,}511e^{-03}$	$6{,}972e^{-31}$	$3{,}380e^{-02}$	$6{,}619e^{-05}$	0,276
$(b_{kt}, \text{électricité}_{kt})$					
Cluster	**1**	**2**	**3**	**4**	**5**
r	0,361	0,041	0,167	0,040	0,345
p.value	$8{,}893e^{-12}$	0,448	$2{,}192e^{-03}$	0,462	$7{,}698e^{-11}$
$(\Delta(b_{kt}), \text{électricité}_{kt})$					
Cluster	**1**	**2**	**3**	**4**	**5**
r	0,494	0,453	0,223	-0,187	0,059
p.value	$5{,}726e^{-22}$	$2{,}151e^{-18}$	$3{,}726e^{-05}$	$5{,}831e^{-04}$	0,283

tections de mouvement dans ces groupes. En outre, comme le graphique 5.21 montre un lien entre les variations des profils de classes et le nombre de détections de mouvement, le

tableau 5.3 calcule également ces corrélations. On observe une corrélation plus élevée entre la dynamique des classes différenciée et le nombre moyen de détections de mouvement. Les corrélations sont calculées également entre la consommation d'électricité moyenne et les profils de classes ainsi qu'avec les profils de classes différenciés. Les corrélations sont plus importantes lorsqu'on considère les variations des profils de classes. Ainsi, il semblerait que ce soit davantage les variations des profils de classes que le niveau qui soit lié à la présence et à la consommation d'électricité.

Le tableau 5.3 fournit également les p-values résultant des tests d'indépendance effectués. Ce test consiste en un test statistique avec l'hypothèse nulle de la nullité de la corrélation. Pour une petite p-value, (< 0.05), cette hypothèse est rejetée et nous pouvons conclure que la corrélation est significativement différente de zéro.

Ces résultats nous permettent d'interpréter les profils de classes comme des modèles d'occupation.

Le cluster 1 présente des pics en fin d'après-midi/début de soirée, ainsi qu'un niveau plus élevé et moins de variation pendant le week-end. Cela indique une présence plus constante et élevée pendant le week-end. Les périodes de présence pendant la semaine sont similaires aux heures de bureau, avec des habitants qui ne travaillent probablement pas le week-end, ce qui expliquerait les variations plus faibles et un niveau plus élevé.

Pour le cluster 2, les pics de présence sont observés le soir (20h/9h) en semaine et ils sont légèrement plus précoces le week-end. De plus, pendant le week-end, les variations sont plus faibles, ce qui indique également une présence plus constante le week-end. En outre, les légers pics du matin indiquent un schéma de présence le matin, mais plus faible et plus court que le soir. Il semble y avoir un léger décalage entre les pics de présence et les pics élevés dans les profils des clusters. Cela peut être dû au temps nécessaire pour que le logement se réchauffe.

Dans le cluster 3, les pics de fréquentation sont assez longs et commencent plus tôt dans la semaine que pour les profils précédents. Cela indique que les personnes sont présentes dès la fin de l'après-midi et pendant la soirée. En outre, il n'y a pas de pic pendant la matinée. Le niveau et les variations sont plus faibles le week-end.

Le profil 4 présente moins de pics périodiques, indiquant une présence moyenne tout au long de la semaine. En outre, le niveau pendant le week-end est plus faible, ce qui pourrait indiquer l'absence des habitants de ces maisons pendant le week-end.

Enfin, le cluster 5 est composé d'une seule maison qui peut être caractérisée comme "atypique". La température intérieure de cette maison présente des variations marquées en début de semaine et beaucoup moins le week-end. De plus, les périodes de présence sont longues et se produisent principalement l'après-midi des jours ouvrables. La comparaison entre le profil des classes et le nombre de détections de mouvements pour ce groupe confirme l'observation faite précédemment.

Les résultats précédemment présentés concernent des données au cours d'une semaine. Cela permet de capter des dynamiques hebdomadaires et définir des semaines types. Cependant, il est également intéressant d'utiliser une plus longue période d'observation et s'intéresser à des dynamiques mensuelles. Les données d'entrée ainsi que les résultats obtenus sur des données d'une période de 4 semaines sont présentés dans l'annexe B. La classification de courbes d'un mois permet d'une part d'identifier des maisons atypiques ayant des périodes d'absences longues et d'autre part des comportements qui peuvent se caractériser par des dynamiques journalières (heures et occurrence des pics) ainsi que hebdomadaires (semaines avec des niveaux plus élevés que d'autres).

Ainsi, pour conclure, l'application du modèle de classification et l'estimation de l'effet des facteurs exogènes ainsi que la dynamique des classes ont permis d'estimer cinq classes parmi 18 maisons individuelles. Les centres des classes ont été comparés aux données des détecteurs de présence. Cette étape a permis d'interpréter les clusters et les profils de classes en matière de comportements de présence au sein des maisons, montrant que chaque cluster peut-être caractérisé par une dynamique d'occupation.

5.3.2 Classification de données de température avec estimation d'effets exogènes propres à chaque cluster

Un second modèle a été présenté dans le chapitre 3 qui cherche à estimer des effets régressifs propres à chaque classe. En effet, dans le cas où les logements ont des isolations ou des expositions différentes, ou encore un mode de chauffage différent, il est possible que les facteurs exogènes tels que la température extérieure ou l'irradiance solaire n'ait pas un effet global et commun.

L'application qui suit concerne les données de la base de données REFIT [Firth et al., 2017] pendant 1 semaine. La sélection des facteurs exogènes est faite à partir du critère BIC comme précédemment et les ensembles candidats sont présentés dans la sous-section 5.2.2. Finalement, 5 classes sont construites et les effets régressifs sont estimés à partir de la température et de l'irradiance. La figure 5.23 représente les températures intérieures et les facteurs exogènes utilisés comme entrées pour l'estimation du modèle. Il s'agit de le même séquence de données utilisée dans l'application présentée dans la section 5.3.1

Le graphique 5.24 représente les profils de classes et les effets régressifs estimés. Pour commencer l'exploration des résultats, la figure 5.24(B) représente les effets exogènes ($\mathbf{a}_k\mathbf{u}_t$) estimés via le modèle. On remarque que l'importance des facteurs exogènes sur la température intérieure est plus ou moins élevée d'une classe à l'autre. On peut notamment noter que les maisons au sein de la classe 1 semblent moins impactées par les conditions météorologiques extérieures que les maisons qui constituent les classes 3 et 5 par exemple. Ces différences peuvent venir d'une exposition différente des pièces à vivre des maisons ou encore d'une isolation plus ou moins performante.

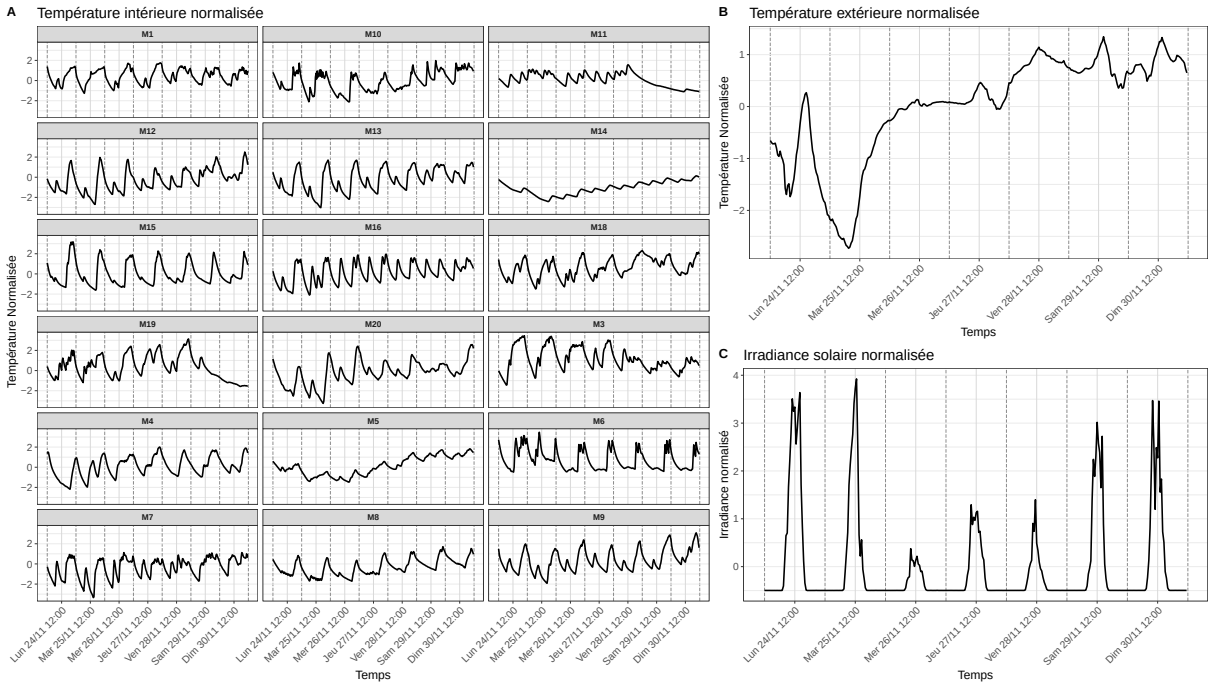

FIGURE 5.23 – Données d'entrée utilisé pour l'application du modèle de classification avec effets régressifs propres à chaque cluster. (A) représente les données de température intérieure normalisées pour 18 maisons anglaises. (B) représente la température extérieure normalisée utilisée comme facteur exogène. (C) représente l'irradiance solaire normalisée utilisée comme facteur exogène.

Ensuite, les profils latents (b_{kt}) des classes, vus comme la part de la température intérieure qui ne peut s'expliquer par des facteurs connus, ont été estimés. La figure 5.24(A) représente les 5 profils obtenus. On peut observer certaines différences entre ces profils, notamment les pics hauts et les pics bas n'apparaissent pas au même moment.

Comme précédemment, les profils de classes sont comparés aux données de détection de mouvement. Comme nous pouvons l'observer dans la figure 5.25, ces deux quantités sont liées. La durée du pic observé sur les profils estimés peut être reliée à la période pendant laquelle le nombre de mouvements observés est élevé. De la même manière, l'heure des pics semble être liée aux heures des pics de présence avec un léger décalage dû à l'inertie de la température par exemple.

Ces observations achèvent l'application des deux modèles sur les données de température de la base de données REFIT. On peut établir un lien entre les profils de classes estimés et la présence des occupants au sein des logements. Cela permet d'interpréter les clusters en fonction des habitudes d'occupation et des emplois du temps des ménages.

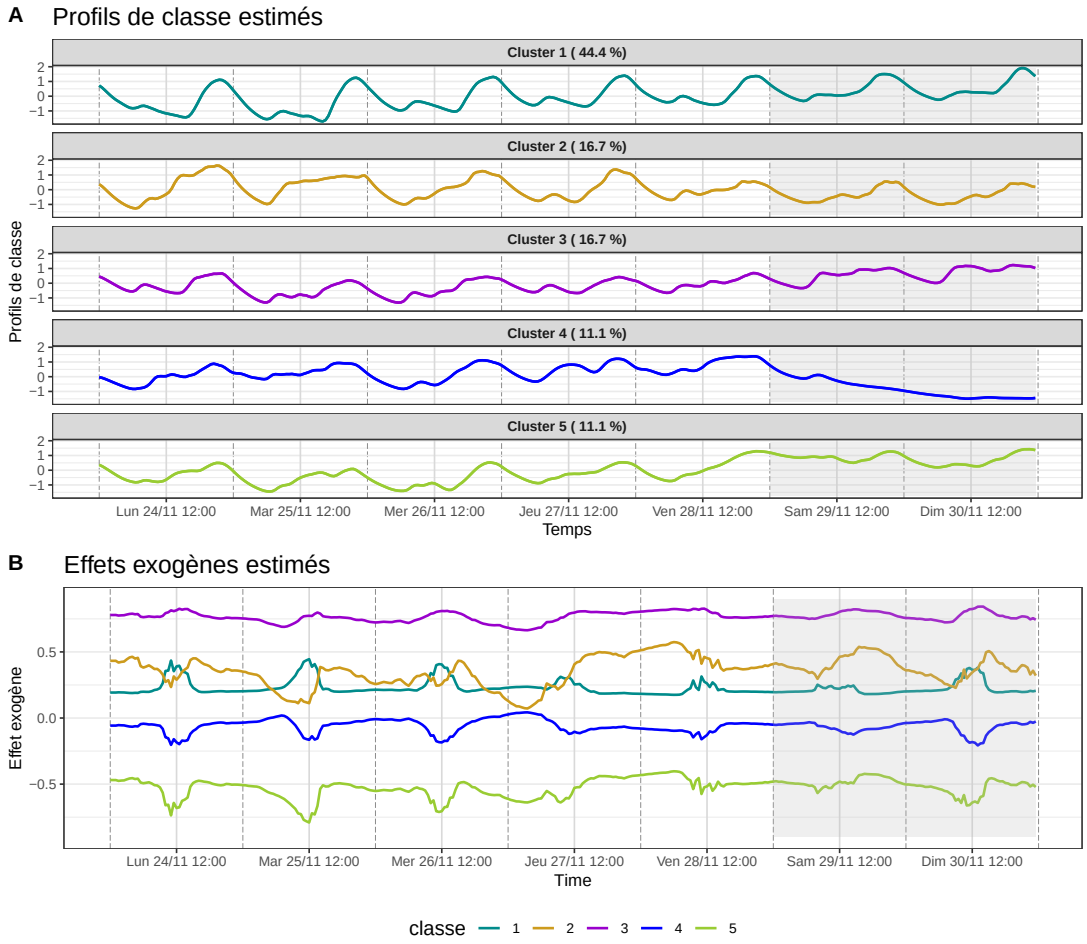

FIGURE 5.24 – Profils de classes estimés et effet régressifs estimés pour chaque classe à partir des données de température de 18 maisons anglaises au cours d'une semaine.

5.4 Classification des données de parties communes de la base de données ANDRE et estimation des profils et des effets exogènes de classes

Le projet ANDRE, présenté dans le chapitre 2 vise à l'analyse de données issues d'un ensemble d'appartements situé dans 3 immeubles en banlieue parisienne. Les données collectées concernent les parties communes des immeubles avec un chauffage collectif où des capteurs de température ont été installés à différents étages. Cette section vise à appliquer le modèle de classification proposé sur ces séquences de température afin de classifier les zones communes et d'estimer des profils de classes et effets exogènes propres à chaque classe.

La figure 5.26 représente les températures mesurées pendant une semaine par les neuf capteurs, ainsi que les données de température et d'humidité extérieure qui seront considérées comme des variables exogènes. Le détail de l'emplacement des capteurs est donné dans le tableau 2.1.

Le modèle de classification avec estimation des effets exogènes propres à chaque classe est

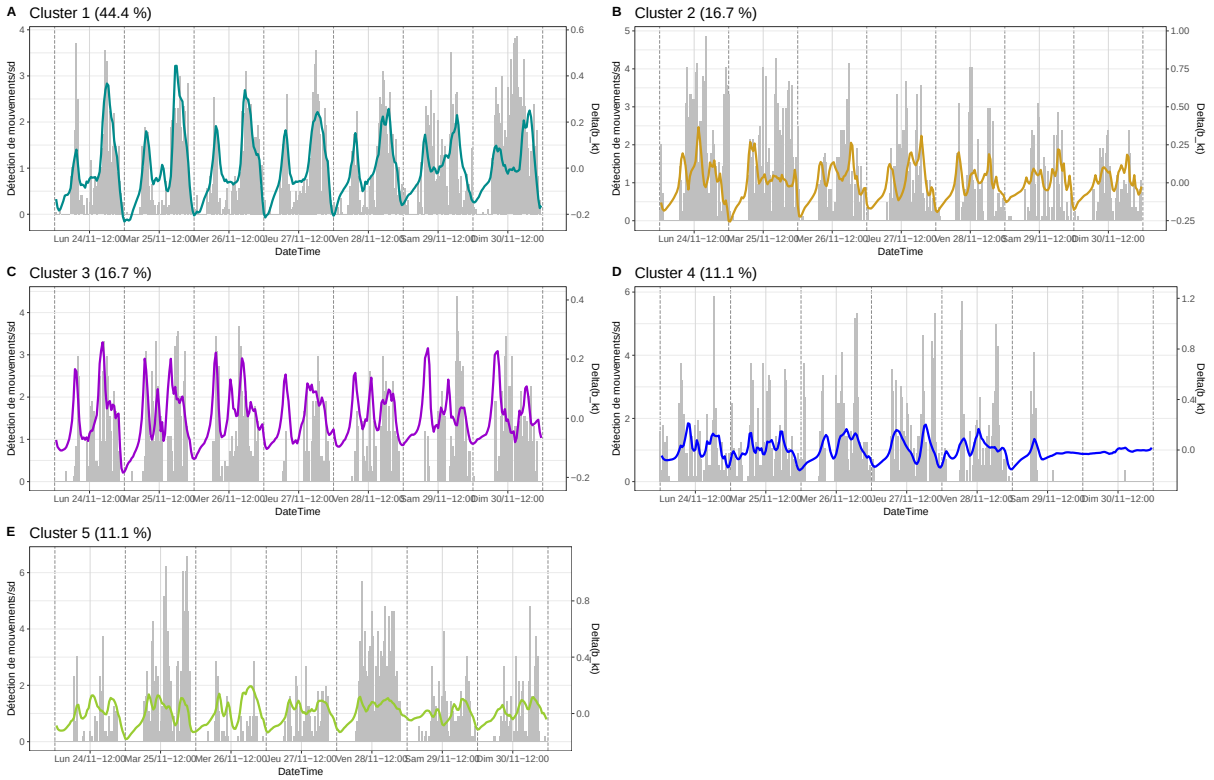

FIGURE 5.25 – Superposition du nombre de détections de mouvements et des profils de classes estimés et différenciés à partir du modèle de classification avec estimation des effets régressifs propres à chaque classe.

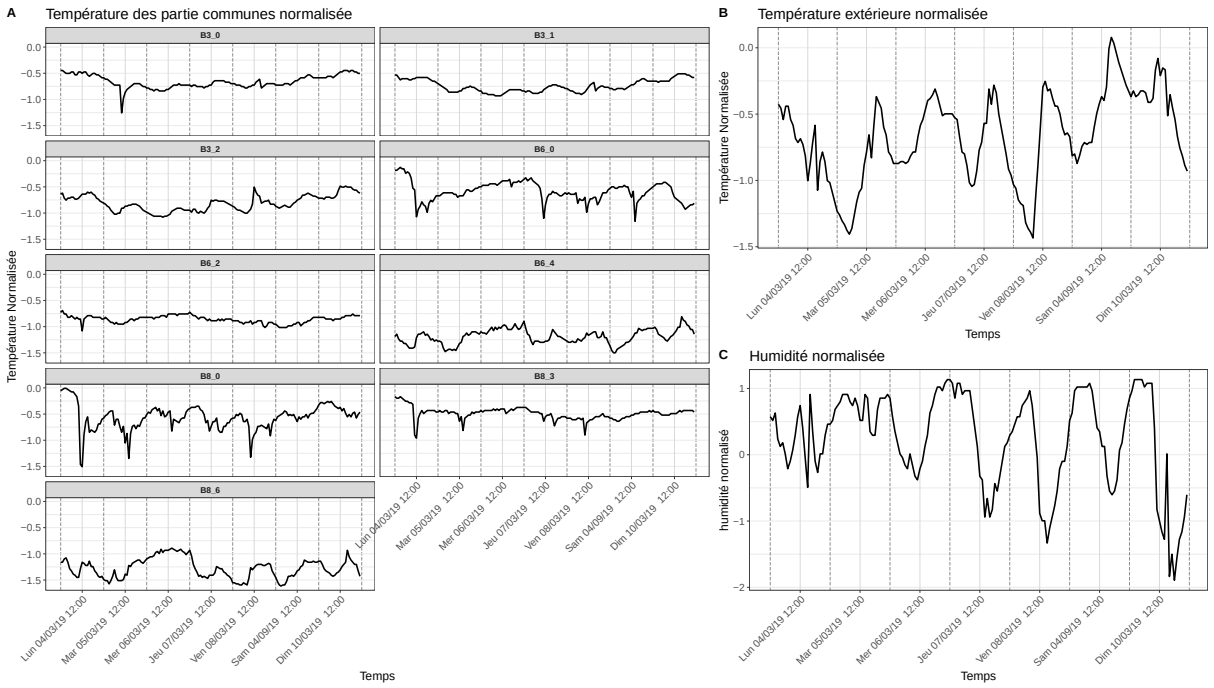

FIGURE 5.26 – Données d'entrée de la base de donnée ANDRE. (A) représente les données de température normalisées mesurées par les neufs capteurs. (B) et (C) représentent respectivement la température extérieure normalisée et l'humidité normalisée considérées comme facteurs exogènes.

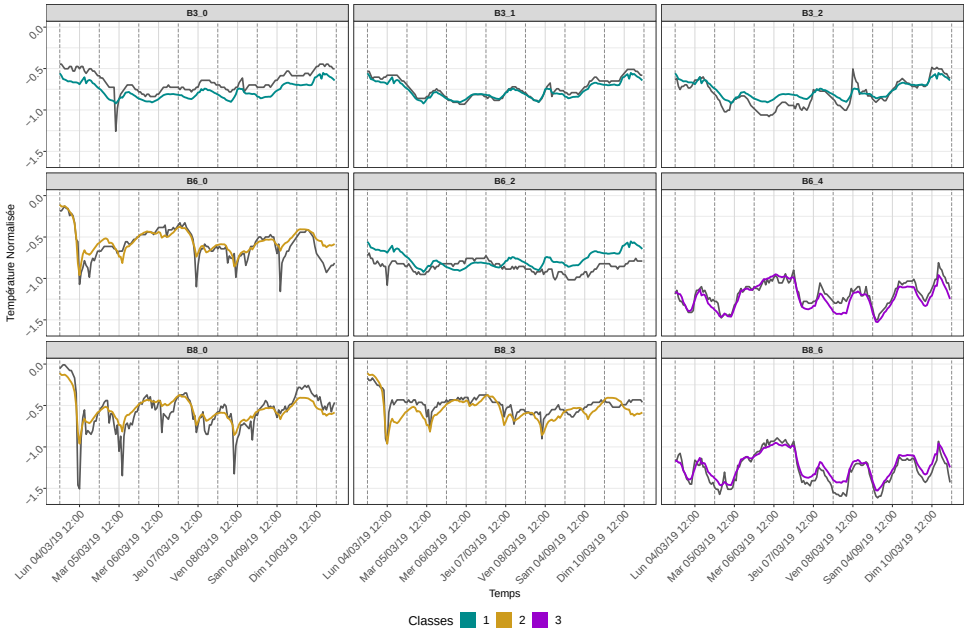

FIGURE 5.27 – Données d'entrée et données estimée selon la classe d'appartenance.

utilisé sur ces données. Les effets sont estimés pour chaque cluster. Les capteurs étant placés à différents niveaux des bâtiments de tailles différentes, on souhaiterait mettre en évidence des sensibilités à la météo extérieure propres à chacun. Ainsi, la température et l'humidité extérieures sont utilisées comme facteurs exogènes et trois clusters sont construits. En effet, on décide de construire trois cluster car dans ce cas, le critère du BIC ne permet pas de définir un nombre de classe précis. De plus, au-delà de trois classes construites, les nouvelles classes ne contiennent qu'un capteur. Pour finir, le choix de construire trois classes est motivé par le fait que les capteurs sont installés dans trois bâtiments à trois niveaux différents. On pourrait donc imaginer que les clusters soient liés au bâtiment d'appartenance ou à l'emplacement des capteurs.

Ici, le nombre d'observation est de 9. Il s'agit d'un petit ensemble de données qui limite l'analyse des résultats. On pourrait imaginer des méthodes artificielles pour augmenter le nombre d'observation. La première possibilité est d'utiliser des méthodes de réechantillonnage comme la *bagging*. Une autre possibilité est d'utiliser plusieurs séquences d'une semaine pour chaque capteur en les considérant comme indépendantes. Ainsi, si on considère 4 semaines de données pour les neufs capteurs, le nouveau jeu de données contient alors 36 observations. Dans ce cas, les facteurs exogènes comme les conditions météorologiques extérieures ne sont plus identiques pour l'ensemble des données. Des essais, allant dans ce sens ont été réalisés. Cependant, les profils de classes estimés, quasiment constants, ne permettaient de discriminer les données qu'en termes de niveau global de température. Ces résultats sont probablement causés par le fait qu'on essaye d'estimer un effet exogène commun pour des données mesurées au cours de différentes saisons et semaines. Il s'agit de piste intéressante à poursuivre en intégrant une variable saisonnière par exemple.

La figure 5.27 représente les données de température et les estimations obtenues via le modèle selon la classe d'appartenance. On peut noter que la classe 3 regroupe les deux capteurs situés les plus en hauteur (4^e et 6^e étage) pour lesquels les températures se caractérisent par des variations assez importantes. Le premier cluster est composé de tous les capteurs du bâtiment 3 ainsi que d'un capteur du bâtiment 6. Les données sont caractérisées par de faibles variations au cours de la semaine. Le deuxième cluster, composé de deux capteurs du bâtiment 8 et d'un capteur du bâtiment 6 se caractérisent par températures qui chutent et remontent rapidement.

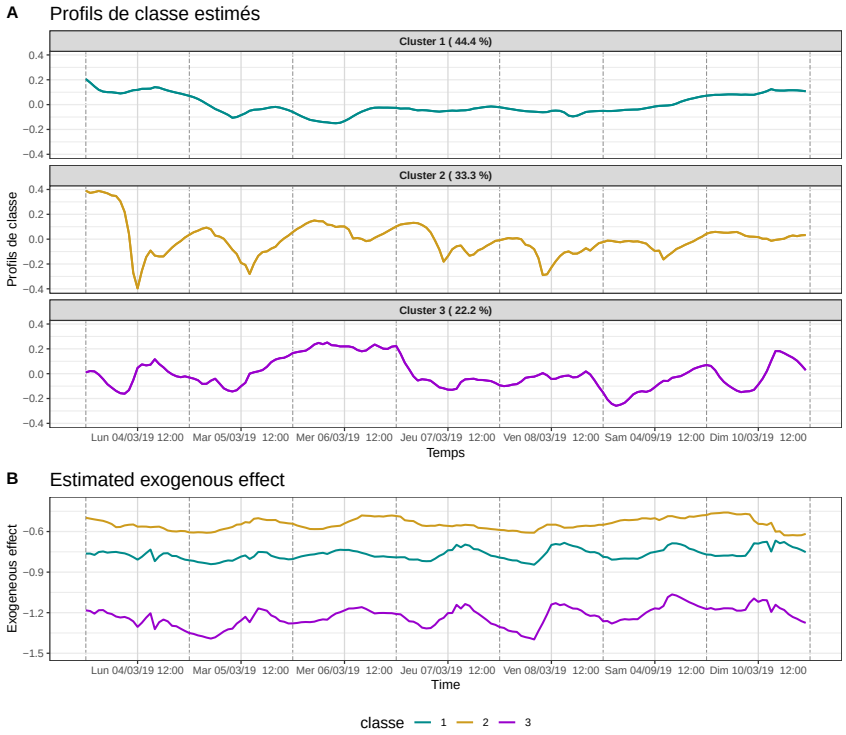

FIGURE 5.28 – Profils de classes et effets exogènes estimés pour les données ANDRE. (A) représente les profils de classes estimés et (B) représente les effets régressifs estimés à partir de la température et de l'humidité extérieure.

TABLE 5.4 – Coefficients estimés pour les 3 clusters.

	Biais	Température	Humidité
Cluster 1	-0.71	0.07	-0.04
Cluster 2	-0.47	0.125	0.05
Cluster 3	-1.09	0.21	-0.01

La figure 5.28 représente les profils de classes (5.28(A)) et les effets exogènes estimés (5.28(B)). On peut noter que le premier cluster est caractérisé par un profil de classe très lisse, à l'image des données de températures. L'effet exogène est, lui aussi, assez lisse bien qu'on observe des variations au cours de la fin de semaine. Le second cluster est caractérisé par un niveau de l'effet exogène plus élevé que les deux autres. Le profil de classe se caractérise par

des pics bas et courts au milieu de la journée. Il s'agit du seul cluster parmi les trois où l'on retrouve une certaine périodicité dans le profil de classe. Le troisième cluster est caractérisé par un effet exogène globalement plus faible, car les données de température mesurées par les capteurs sont plus faibles (cf 5.27). Le tableau 5.4 présente les coefficients estimés pour chacun des trois clusters. On peut noter que le cluster 3 est celui pour lequel la température extérieure à l'effet le plus important, alors que l'humidité à un effet négatif et faible. Le cluster 2 est impacté positivement par la température et par l'humidité. Pour finir, le premier cluster présente le coefficient le plus faible pour la température extérieure, ce qui peut expliquer les faibles variations des températures mesurées.

5.5 Conclusion

Ce chapitre a été consacré à l'application des deux modèles construits sur différents jeux de données thermiques. Dans un premier temps, l'utilisation d'un modèle de simulation thermique a permis de construire un jeu de données de température pour un ensemble d'appartements à partir de signaux de thermostat type basés sur les températures de consigne. La classification de ces données permet d'estimer de profils de classes qui correspondent aux consignes de température initialement utilisées. Cela permet d'interpréter les classes en termes d'habitude de chauffage. Dans un second temps, le modèle a été appliqué à des séquences individuelles afin d'explorer sa capacité à séparer les effets exogènes liés à des facteurs mesurables des effets endogènes propres aux comportements des habitants. Cette application permet, *a priori*, de séparer les effets périodiques liés à l'heure de la journée via l'effet régressif et d'estimer, via le profil de classe, les variations des comportements au cours de la semaine.

Dans un second temps, les deux modèles proposés sont appliqués à des données de température de 18 maisons anglaises issues de la base de données REFIT. L'application, sur des données d'une semaine, puis d'un mois, du modèle de classification avec estimation d'un effet régressif commun permet d'estimer de profils de classes qui peuvent être interprétés comme des profils de présence. En effet, le croisement des profils de classes aux données de détection de mouvement et de consommation d'électricité permet d'interpréter ces résultats comme des comportements d'occupation, et de confirmer le lien entre la température intérieure et l'occupation des logements. Ensuite, ces données sont utilisées pour la classification et l'estimation d'effets régressifs propres à chaque cluster, à l'aide du second modèle. Les résultats obtenus sont également liés à la présence des occupants dans les logements, mais permettent aussi de mettre en évidence différentes sensibilités aux conditions météorologiques extérieures.

Pour finir, les données de température de parties communes de bâtiments, issues de la base de données ANDRE, sont classifiées à l'aide du modèle avec effets exogènes propres à chaque cluster. Cette application permet de différencier les capteurs pour lesquels les

températures sont très variables de ceux où les températures sont plutôt constantes au cours de la semaine. De plus, il semblerait que les capteurs, selon leur emplacement, soient plus ou moins influencé par les conditions météorologiques extérieures.

Pour conclure ce chapitre, les deux modèles de classification, appliqués à des données de température au sein de logements, permettent de construire des clusters dont les profils de classes peuvent être interprétés comme des comportements de chauffage ou d'occupation.

Chapitre 6

Conclusion et perspectives

Ces travaux se placent dans le cadre de la classification de données énergétiques et thermiques dans le but d'extraire des dynamiques de comportements type d'habitants. Pour cela, on a proposé un premier modèle de classification avec des centres de classes dynamiques et un effet régressif commun. Ce modèle permet de regrouper des données temporelles sous forme de clusters homogènes tout en modélisant les centres de classes sous la forme de processus stochastique dans le but de prendre en compte la dynamique des comportements au cours du temps. Le modèle cherche également à estimer un effet régressif commun à l'ensemble des observations relatif à des variables exogènes observables. Ce modèle a ensuite été évalué et comparé à des modèles de référence, à partir de critères d'évaluation et d'un ensemble de jeux de données simulées. Cette comparaison a permis de mettre en évidence que le modèle proposé offre de meilleures performances sur la base des trois critères choisis et des jeux de données proposés. De plus, cette évaluation confirme également que le modèle est d'autant plus performant que le nombre de données est important. L'extension du premier modèle, qui consiste à estimer les effets régressifs pour chaque cluster, a été présentée. La comparaison des deux modèles proposés a permis de mettre en évidence que le premier modèle est préférable dans le cas où il ne serait pas possible de savoir, *a priori,* si les données sont sujettes à des effets régressifs différents selon le cluster d'appartenance.

Ensuite, la simulation d'un jeu de données à partir d'un modèle de simulation thermique a permis de construire une base de données de température pour 15 appartements afin d'évaluer le modèle dans un cadre réaliste. L'application du modèle de classification permet d'identifier les classes et les profils de classes qui sont relatifs aux thermostats utilisés pour la simulation. Dans un second temps, le modèle de classification proposé est appliqué à des séquences individuelles de température pour un seul appartement dans le but de séparer les effets liés à des facteurs exogènes de ceux relatifs au comportement endogène des habitants.

Par la suite, la base de données REFIT, regroupant des données de maisons individuelles anglaises, a été utilisée pour la classification des données de température. Les profils de classe estimés ont été mis en relation avec le nombre de détections de mouvement. Ces résultats ont permis d'interpréter les profils de classes comme des comportements d'occupation type

et d'interpréter les classes selon les emplois du temps des ménages. Pour finir l'application des modèles à des données réelles, la base de données ANDRE, qui regroupe des données de partie communes issues de 3 bâtiments, est utilisée pour la classification à l'aide du modèle avec centres dynamiques et effets régressifs propres à chaque classe. Les résultats ont mis en évidence des différences de sensibilité aux conditions météorologiques extérieures selon l'emplacement des capteurs.

Cette thèse a été initiée par le projet ANDRE évoqué précédemment. Ce projet prévoyait la collecte de données au sein d'un ensemble d'appartements. L'une des principales perspectives est d'appliquer les modèles présentés dans ces travaux aux données de températures mesurées au sein des appartements pour en extraire des comportements d'occupants.

Ensuite, nous avons évoqué que l'un des aspects intéressant dans le choix de processus autorégressifs pour modéliser les profils de classe résidait dans la possibilité d'utiliser les profils estimés pour prédire des valeurs futures. L'aspect prédictif de ce modèle n'a pas été approfondi dans ces travaux, mais pourrait constituer une perspective intéressante. De plus, l'application sur des données réelles de température intérieure permet d'interpréter les profils de classe selon des habitudes d'occupation. Dans des travaux futurs, la dynamique estimée des classes pourrait être utilisée pour estimer l'occupation au sein des logements.

Enfin, le choix de modéliser les profils de classe à partir d'un processus autorégressif de premier ordre est un sujet de discussion. On peut imaginer une étape supplémentaire afin de sélectionner un ordre supérieur qui serait mieux adapté aux données. Cela nécessite soit de sélectionner cet hyperparamètre dans une étape précédente, soit de développer une méthode pour intégrer l'ordre des processus comme paramètres du modèle.

Annexe A

Description simplifiée du modèle thermique

Le modèle thermique a pour objectif de déterminer la température de la zone (température intérieure) et celle de la paroi à partir de consigne de température, de la température extérieure, de l'irradiance solaire et d'un ensemble de paramètre caractérisant la zone d'intérêt. Dans un premier temps, il faut définir un ensemble de paramètres et de données d'entrée au modèle de simulation. Ensuite, une suite d'équation permet de calculer les températures et la puissance des radiateurs.

A.1 Les paramètres et données d'entrée

Afin de simuler des températures intérieures et de paroi, il faut d'abord définir des paramètres et données d'entrée. On dispose des éléments suivants :

- Le nombre de pas de temps pour la simulation $n_t = 8928$. Cette dernière, dans notre cas, est réalisée sur une période de 31 jours au pas de temps 5 minutes. De cela découle le paramètre $d_t = \frac{\tau}{n_t}$, où τ correspond au temps d'étude, soit τ = 31x24x3600 secondes.
- Le nombre d'éléments discrétisés $n_d = 46$. La paroi en contact avec l'extérieur de la zone est discrétisée sous forme de 46 éléments. Ainsi, le modèle thermique va chercher à simuler la température des 46 points de la paroi ainsi qu'un point correspondant aux échanges avec l'extérieur et un autre pour la zone intérieure, soit 48 points au total.
- La surface de la paroi opaque (mur) est notée S_e et la surface vitrée est notée S_v.
- Les vecteurs de température extérieure et d'irradiance solaire indirecte notés respectivement $\mathbf{y}_1$ et $\mathbf{y}_2$. On associe deux coefficients β_1 et β_2 à l'irradiance solaire. Ces coefficients permettent de mesurer l'apport du rayonnement via le mur (β_1) et via la fenêtre (β_2). Ils sont dimensionnés selon le type de fenêtre et le type de paroi et leur surface. On a $\mathbf{y}_1 \in \mathbb{R}^{nd}$ et $\mathbf{y}_2 \in \mathbb{R}^{nd}$.

- Les vecteurs de température de consigne et de ventilation, notés respectivement **tc** et **v**, avec $\mathbf{tc} \in \mathbb{R}^{nd}$ et $\mathbf{v} \in \mathbb{R}^{nd}$.
- Le coefficient, noté c_a, correspond à la capacité thermique volumique de l'air. Le coefficient noté c correspond à la capacité volumique de la zone en tenant compte de la présence du mobilier.
- Le coefficient d'échange entre l'enveloppe intérieure et la zone est noté h_i et celui entre l'enveloppe extérieure et l'extérieur est noté h_e.
- Une description de l'enveloppe et des couches qui la compose. Chaque couche est décrite par son épaisseur, son coefficient de conductivité et de capacité thermique. Ces données sont décrites dans le tableau 5.1.

A.2 Simulation des vecteurs de température de la zone, de la paroi et puissance des radiateurs :

La simulation consiste à déterminer d'une part la température des 47 points de la paroi ainsi que celle de la zone. On cherche également à déterminer la puissance des radiateurs. Pour cela, la résolution consiste, pour chaque pas de temps, à la répétition de plusieurs étapes. On définit d'abord les éléments suivants :

$\mathbb{C}$ **:** Une matrice construite à partir de la capacité thermique de chaque couche constituant la paroi, l'épaisseur de ces couches et la capacité thermique de la zone (voir le tableau 5.1).

$\mathbb{K}$ **:** Une matrice construite à partir des coefficients de conductivité thermique des couches de la paroi, le coefficient d'échange thermique entre la zone et la paroi intérieure et celui entre la paroi extérieure et l'extérieur (voir le tableau 5.1).

$\mathbb{A}_1$ **:** Une matrice construite à partir des matrices $\mathbb{K}$ et $\mathbb{C}$ qui permet de relater la conductivité et la capacité thermique de chaque élément de la paroi et de la zone, sans prendre en compte la ventilation. Avec $\mathbb{A}_1 \in \mathscr{M}_{(48,48)}$.

$\mathbb{A}_2$ **:** avec $\mathbb{A}_2 \in \mathscr{M}_{(48,48)}$.

W_0 **:** La puissance du chauffage qui dépend de la taille de la pièce.

$\mathbf{t}_z^{(t)}$ **:** Le vecteur de la température simulée au pas de temps t, avec $\mathbf{t}_z = (t_{z1}, ..., t_{z48})$.

h : le vecteur de puissance de radiateur, avec $\mathbf{h}\ \mathbb{R}^{n_t}$.

$\mathbf{f}_t$ **:** Le vecteur des changements thermique au temps t. Avec $\mathbf{f}_t = (f_{t1}, ..., f_{t48}) \in \mathbb{R}^{48}$. Ces vecteurs sont initialisés avec des 0.

ϵ **:** la tolérance acceptée entre la température de la pièce et celle de consigne avant de déclencher ou arrêter le chauffage.

La résolution du problème de simulation consiste en la répétition des étapes suivantes pour chaque pas de temps t :

1. Apport du chauffage :

$$h_t = \begin{cases} W_0, & \text{if } \mathbf{t}_{z1}^{(t-1)} - \epsilon < \mathbf{tc}_{t-1}. \\ 0, & \text{sinon.} \end{cases} \tag{A.1}$$

2. Calcul du premier élément du changement thermique :

$$f_{t1} = c_a y_{1t} v_t + h_t + \beta_1 y_{2t} \tag{A.2}$$

$$f_{t48} = h_e S_e y_{1t} + \beta_2 y_{2t} \tag{A.3}$$

3. Calcul du second membre pour le changement thermique :

$$\tilde{\mathbf{f}}_t = \frac{1}{dt}\mathbb{C}\mathbf{t}_z^{(t-1)} + \mathbf{f}_t \tag{A.4}$$

4. Mise à jour de la matrice $\mathbb{A}_2$ avec la ventilation :

$$\mathbb{A}_2 = \mathbb{A}_1 \tag{A.5}$$

$$\mathbb{A}_2(1,1) = \mathbb{A}_2(1,1) + c_a v_t \tag{A.6}$$

5. Calcul de la température des éléments de la paroi et de la zone :

$$\mathbf{t}_z^{(t)} = \mathbb{A}_2 \mathbf{f}_t^{-1} \tag{A.7}$$

Ainsi, la température de la zone au temps t est donnée par $t_{z1}^{(t)}$ et les températures des éléments de la paroi par le vecteur $(t_{z2}^{(t)}, ..., t_{z48}^{(t)})$. La puissance des radiateurs est donnée par le vecteur $\mathbf{v}$.

Annexe B

Application du modèle de classification avec centres dynamiques et effet exogène commun sur des données d'un mois

La figure B.1 représente les mesures de températures centrées et réduites pour l'ensemble des 18 maisons entre le 24 novembre et le 21 décembre 2014, soit 4 semaines complètes.

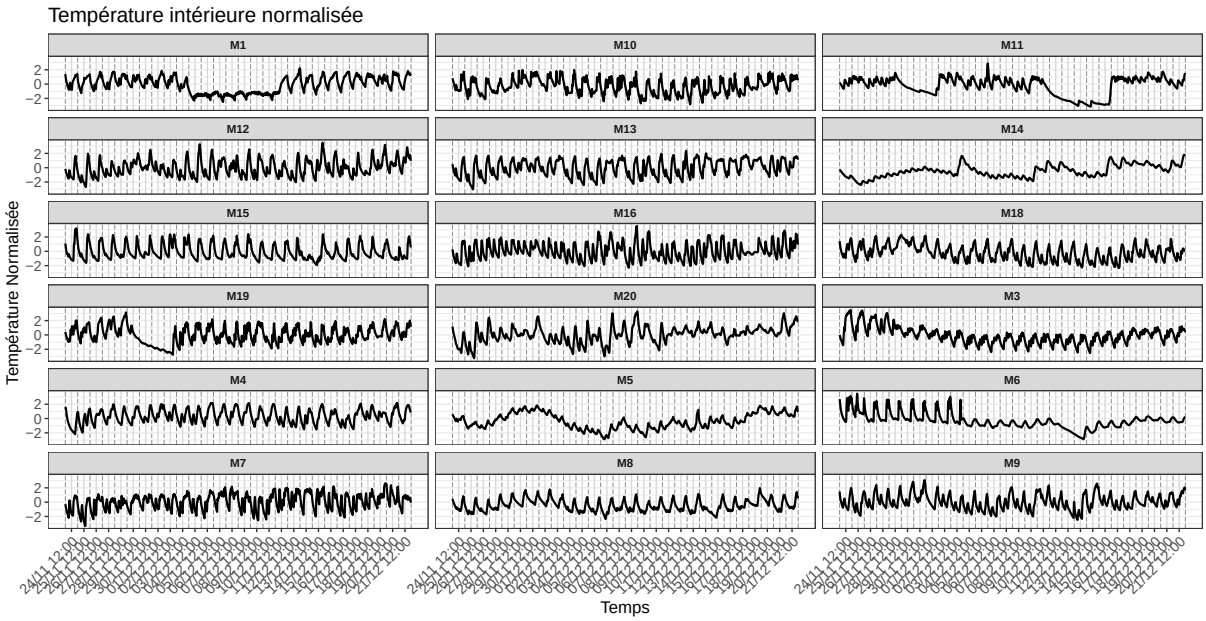

FIGURE B.1 – Températures intérieures normalisées mesurées au sein de 18 maisons au cours de la semaine du 24 novembre au 21 décembre 2014.

Les facteurs exogènes utilisés, présentés dans la figure B.2, ont été sélectionné comme précédemment avec le critère BIC. Ainsi, seules la température extérieure et l'irradiance solaire sont utilisées comme facteurs exogènes communs après avoir été normalisés. Ensuite, 4 classes sont construites. En effet, au-delà, les nouvelles classes construites ne contiennent qu'une seule maison.

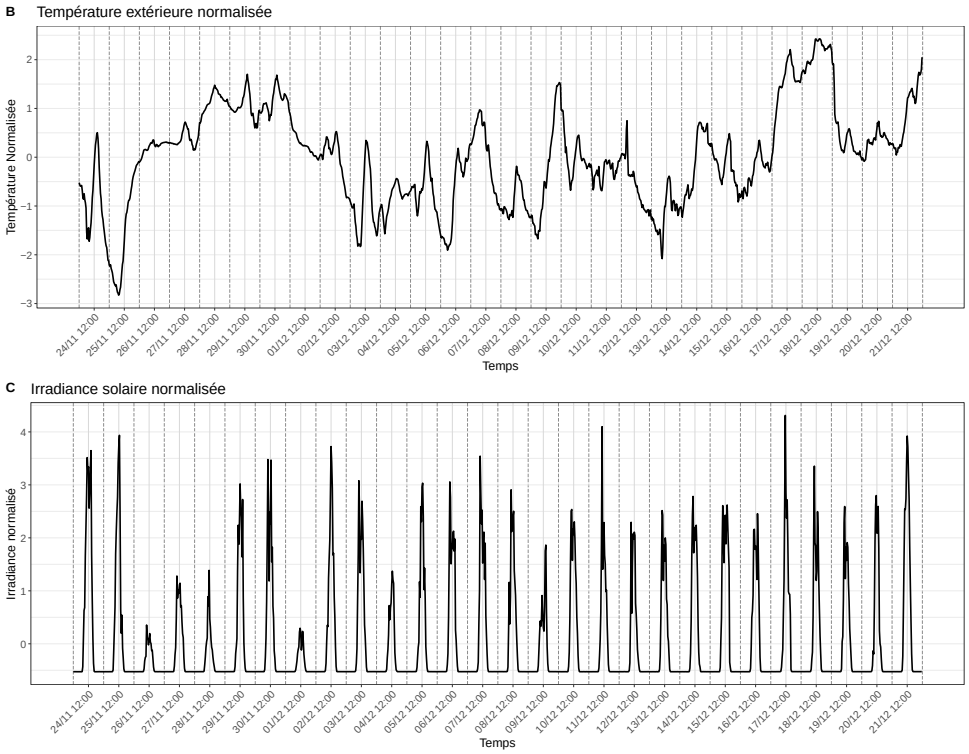

FIGURE B.2 – La température extérieure normalisée (A), l'irradiance solaire normalisée (B) pendant 4 semaines, au pas de temps 30 minutes utilisées comme facteur exogènes.

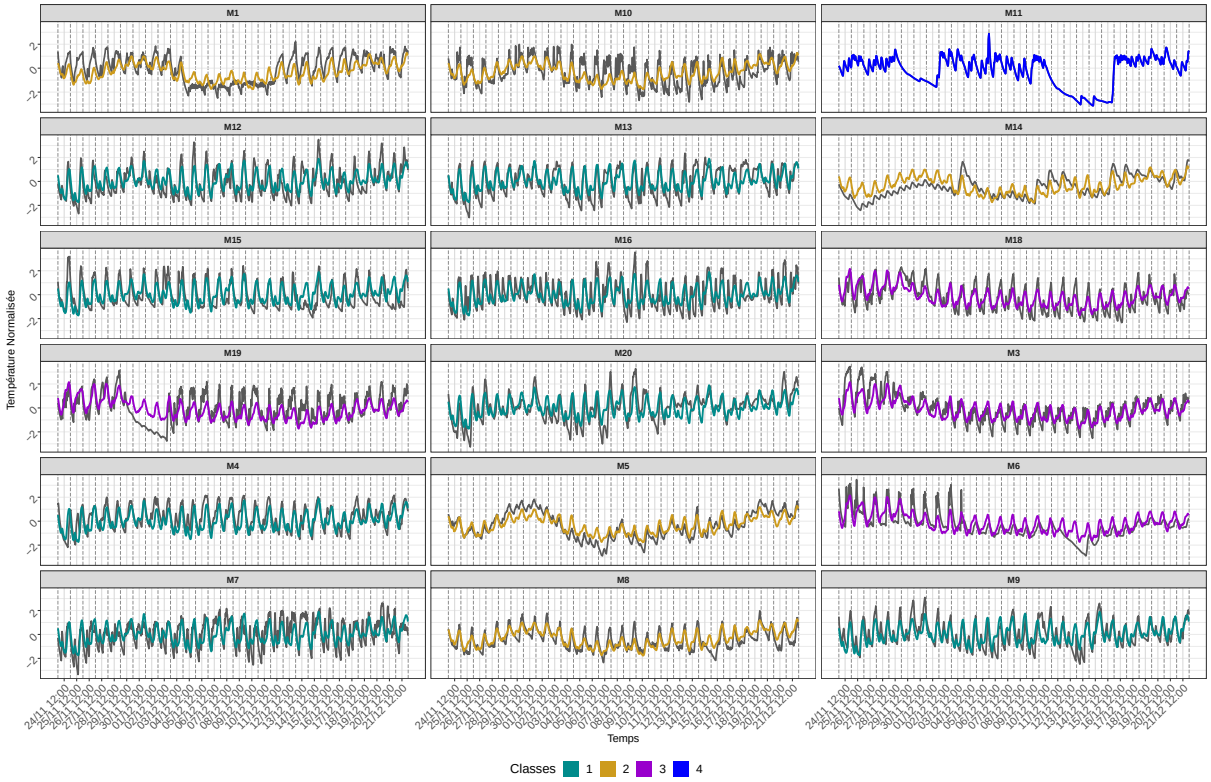

FIGURE B.3 – Données de température pour 18 maisons anglaises au cours de 4 semaines et les données estimées via le modèle de classification selon la classe d'appartenance estimée.

La figure B.3 représente les données de température mesurées et estimées via le modèle ainsi que les classes construites. Les estimations ont été obtenues via la formule (5.3). On peut noter que la classe 1 est majoritaire et regroupe 8 des 18 maisons. Au contraire, la classe 4 ne contient qu'une seule maison. Ce résultat se justifie par la présence de deux périodes atypiques dans les mesures de température au sein de la maison 11.

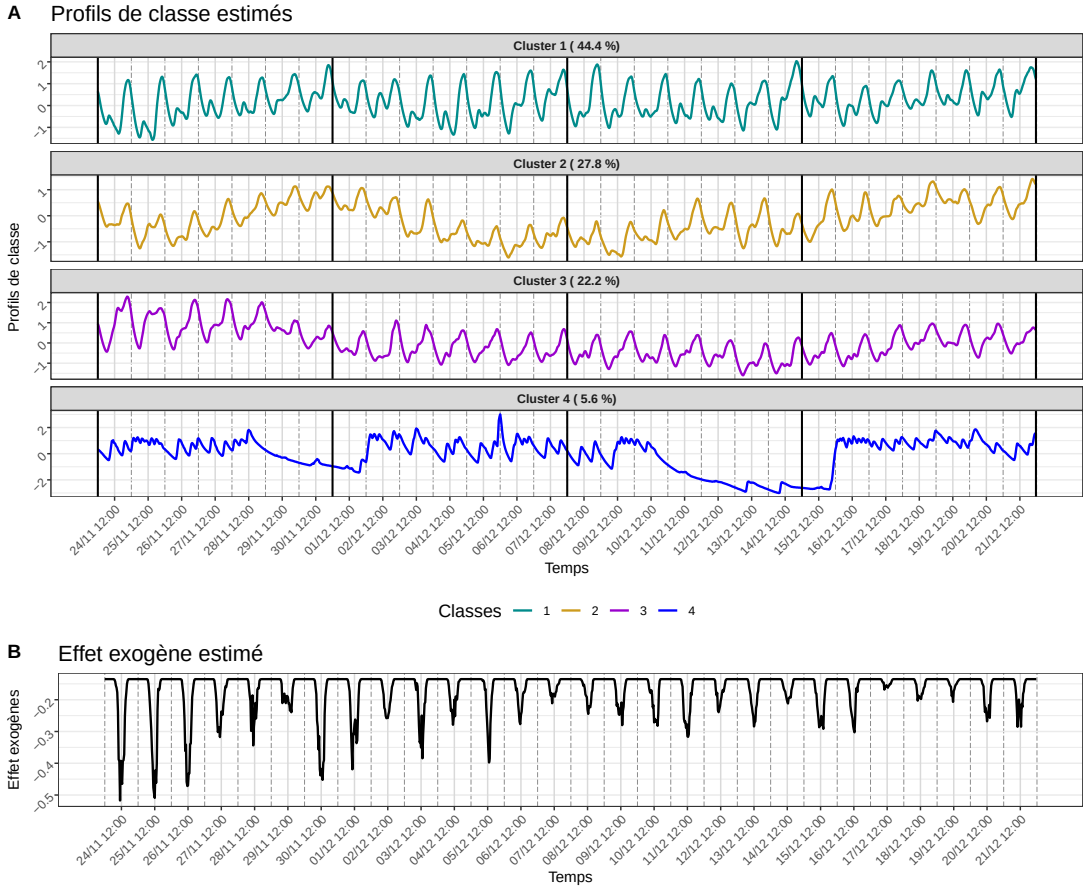

FIGURE B.4 – Profils de classes et effet exogènes estimés via le modèle à partir des données de 18 maisons anglaises pendant 4 semaines.

La figure B.4 représente les profils de classes estimés et les effets des facteurs exogènes. Le premier profil, le plus majoritaire, se distingue d'une part par deux pics journaliers qui se répètent au cours des 4 semaines. Le schéma hebdomadaire au cours des 4 semaines est assez répétitif avec un niveau plus élevé les des week-end et des pics plus importants. Les profils de classes se distinguent par l'occurrence des pics au cours de la journée, les dynamiques différentes entre la semaine et les week-end et des variations d'une semaine à l'autre. La figure B.4 présente l'effet exogène estimés liés à la température extérieure et à l'irradiance solaire. On peut observer des pics bas, aux alentours de midi. Ici considérer plusieurs semaine pourrait permettre de mettre en évidence des classes de maison ayant des périodes d'absences marquées, comme c'est le cas pour la maison n°11.

Afin, comme précédemment, d'interpréter les résultats, on peut mettre en relation les centres de classes estimés et les données de détection de mouvements au sein de maisons

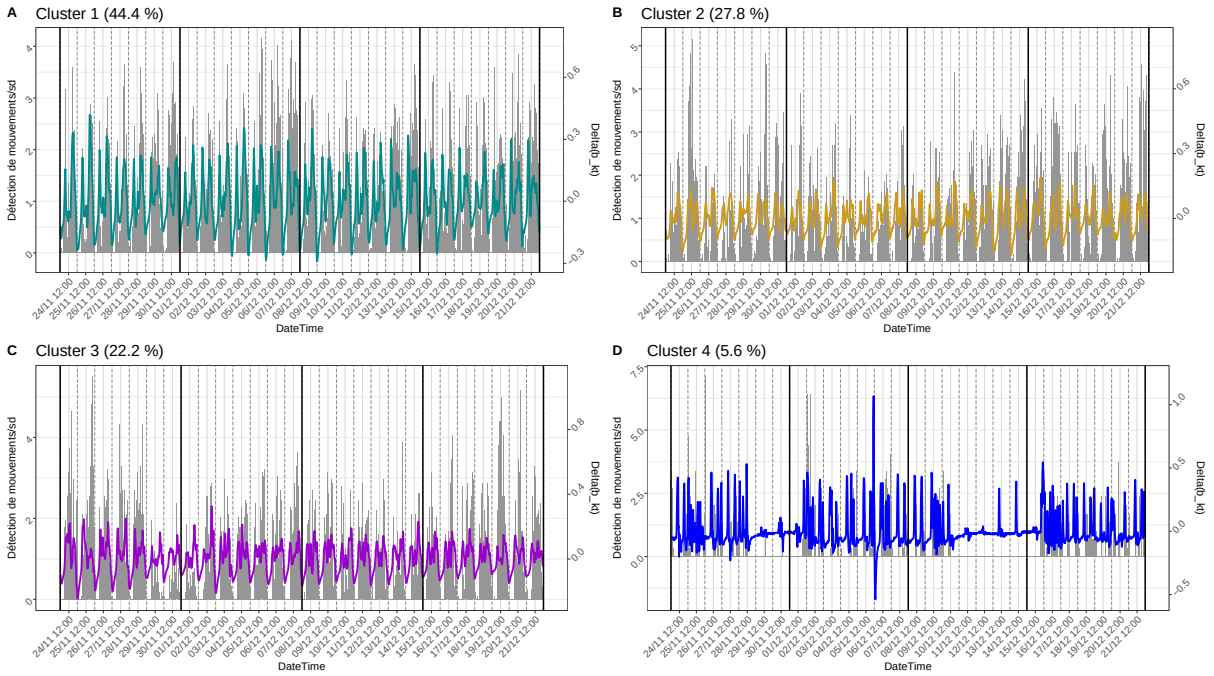

FIGURE B.5 – Superposition des profils de classes différenciés et majoritaires avec le nombre de détections de mouvement moyen au cours des 4 semaines.

sur la période. Ces résultats sont représentés sur le graphique B.5. Un lien peut être établi entre les profils de classes et le nombre de mouvements détecté. En effet, par exemple, pour le profil 2, l'augmentation du niveau de processus estimé peut-être justifiée par une présence plus importante au cours des deux dernières semaines. Les périodes de présence semblent plus longues et le niveau global plus important, notamment au cours de la dernière semaine. Concernant le 3^{e} profil, on avait observé une amplitude des pics et un niveau global plus important au cours de la première semaine. On observe une présence plus importante au cours de cette même semaine.

Bibliographie

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6) :716–723.

Albert, A. and Rajagopal, R. (2013). Smart meter driven segmentation : What your consumption says about you. *IEEE Transactions on Power Systems*, 28(4) :4019–4030.

Amayri, M., Ploix, S., Kazmi, H., Ngo, Q.-D., and El Safadi, A. (2019). Estimating occupancy from measurements and knowledge using the bayesian network for energy management. *Journal of Sensors*, 2019 :1–12.

Arlot, S. (2019). Minimal penalties and the slope heuristics : a survey. *Journal de la Societe Française de Statistique*, 160(3) :1–106.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3) :803–821.

Basford, K. E. and McLachlan, G. J. (1985). The mixture method of clustering applied to three-way data. *Journal of Classification*, 2 :109–125.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics : Overview and implementation. *Stat Comput*, 22.

Beckel, C., Sadamori, L., and Santini, S. (2012). Towards automatic classification of private households using electricity consumption data. *BuildSys 2012 - Proceedings of the 4th ACM Workshop on Embedded Systems for Energy Efficiency in Buildings*, pages 169–176.

Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. Technical Report RR-3521, INRIA.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3) :561–575.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1-2) :33–73.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference : A review for statisticians. *Journal of the American Statistical Association*, 112(518) :859–877.

Bourdeau, M., Basset, P., Beauchêne, S., Da Silva, D., Guiot, T., Werner, D., and Nefzaoui, E. (2021). Classification of daily electric load profiles of non-residential buildings. *Energy and Buildings*, 233 :110670.

Burnham, K. and Anderson, D. (2004). Model selection and multimodel inference. *A Practical Information-theoretic Approach*.

Calabrese, A. and Paninski, L. (2011). Kalman filter mixture model for spike sorting of non-stationary data. *Journal of neuroscience methods*, 196(1) :159—169.

Celeux, G. and Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3) :315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793.

Corduneanu, A. and Bishop, C. (2001). Variational bayesian model selection for mixture distribution. *Artificial Intelligence and Statistics*, 18 :27–34.

De Meester, T., Marique, A.-F., De Herde, A., and Reiter, S. (2013). Impacts of occupant behaviours on residential heating consumption for detached houses in a temperate climate in the northern part of europe. *Energy and Buildings*, 57 :313–323.

De Wilde, P. (2014). The gap between predicted and measured energy performance of buildings : A framework for investigation. *Automation in Construction*, 41 :40–49.

Delff Andersen, P., Iversen, A., Madsen, H., and Rode, C. (2014). Dynamic modeling of presence of occupants using inhomogeneous markov chains. *Energy and Buildings*.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977b). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22.

Desarbo, W. and Cron, W. (1988). A conditional mixture maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5 :249–282.

Desarbo, W. and Wedel, M. (2002). *Mixture Regression Models*, pages 366– 382.

Devijver, E., Goude, Y., and Poggi, J.-M. (2015). Clustering electricity consumers using high-dimensional regression mixture models. *Applied Stochastic Models in Business and Industry*, 36.

Diao, L., Sun, Y., Chen, Z., and Chen, J. (2017). Modeling energy consumption in residential buildings : A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings*, 147 :47–66.

Djatouti, Z., Waeytens, J., Chamoin, L., and Chatellier, P. (2020). Thermal behavior of a two-story concrete building under controlled winter and heat wave scenarios in the sense-city equipment through temperature, flux and energy consumption dataset. *Data in Brief*, 33 :106458.

Djatouti, Z., Waeytens, J., Chamoin, L., and Chatellier, P. (2021). Goal-oriented sensor placement and model updating strategies applied to a real building in the sense-city equipment under controlled winter and heat wave scenarios. *Energy and Buildings*, 231 :110486.

D'Oca, S. and Hong, T. (2015). Occupancy schedules learning process through a data mining framework. *Energy and Buildings*, 88 :395–408.

Eirola, E. and Lendasse, A. (2013). Gaussian mixture models for time series modelling, forecasting, and interpolation. volume 8207, pages 162–173, Berlin, Heidelberg. Springer-Verlag.

El Assaad, H., Samé, A., Govaert, G., and Aknin, P. (2016). A variational expectation–maximization algorithm for temporal data clustering. *Computational Statistics & Data Analysis*, 103 :206–228.

Fazeli, R., Ruth, M., and Davidsdottir, B. (2016). Temperature response functions for residential energy demand – a review of models. *Urban Climate*, 15 :45–59.

Firth, S., Kane, T., Dimitriou, V., Hassan, T., Fouchal, F., Coleman, M., and Webb, L. (2017). Refit smart home dataset.

Franco, A. and Leccese, F. (2020). Measurement of co2 concentration for occupancy estimation in educational buildings with energy efficiency purposes. *Journal of Building Engineering*, 32 :101714.

Gershman, S., Hoffman, M., and Blei, D. (2012). Nonparametric variational inference. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1.

Granell, R., Axon, C. J., and Wallom, D. C. (2015). Clustering disaggregated load profiles using a dirichlet process mixture model. *Energy Conversion and Management*, 92 :507–516.

Hurn, M., Justel, A., and Robert, C. (2000). Estimating mixture of regressions. *Journal of Computational and Graphical Statistics*, 12.

Jones, P. and McLachlan, G. (2008). Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, 34 :233 – 240.

Kane, T., Firth, S., Hassan, T., and Dimitriou, V. (2017). Heating behaviour in english homes : An assessment of indirect calculation methods. *Energy and Buildings*, 148.

Lévy, J.-P., Roudil, N., Flamand, A., and Belaïd, F. (2014). Les déterminants de la consommation énergétique domestique : le projet Energihab. *Flux - Cahiers scientifiques internationaux Réseaux et territoires*, 96 :40–54.

Li, Y., Schofield, E., and Gönen, M. (2019). A tutorial on dirichlet process mixture modeling. *Journal of Mathematical Psychology*, 91 :128–144.

Liisberg, J., Møller, J., Bloem, H., Cipriano, J., Mor, G., and Madsen, H. (2016). Hidden markov models for indirect classification of occupant behaviour. *Sustainable Cities and Society*, 27 :83–98.

Lu, Y., Tian, Z., Peng, P., Niu, J., Li, W., and Zhang, H. (2019). Gmm clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system. *Energy and Buildings*, 190 :49–60.

Lücke, J. and Forster, D. (2019). K-means as a variational em approximation of gaussian mixture models. *Pattern Recognition Letters*, 125 :349–356.

Maugis, C. and Michel, B. (2008). Slope heuristics for variable selection and clustering via Gaussian mixtures. Research Report RR-6550, INRIA.

Maugis, C. and Michel, B. (2011). Data-driven penalty calibration : A case study for gaussian mixture model selection. *ESAIM : Probability and Statistics*, 15 :320–339.

McLachlan, G. and Krishnan, T. (2007). *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley.

McLachlan, G. J. (2015). Mixture models in statistics. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 624–628. Elsevier, Oxford, second edition edition.

McLoughlin, F., Duffy, A., and Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, 141 :190–199.

Melzi, F. N., Same, A., Zayani, M. H., and Oukhellou, L. (2017). A dedicated mixture model for clustering smart meter data : Identification and analysis of electricity consumption behaviors. *Energies*, 10(10).

Pan, S., Wang, X., Wei, Y., Zhang, X., Gal, C., Guangying, R., Yan, D., Shi, Y., Wu, J., Xia, L., Xie, J., and Liu, J. (2017). Cluster analysis for occupant-behavior based electricity load patterns in buildings : A case study in shanghai residences. *Building Simulation*, 10.

Paone, A. and Jean-Philippe, B. (2018). The impact of building occupant behavior on energy efficiency and methods to influence it : A review of the state of the art. *Energies*, 11 :953.

Pardo, A., Meneu, V., and Valor, E. (2002). Temperature and seasonality influences on spanish electricity load. *Energy Economics*, 24(1) :55–70.

Randriamanamihaga, A. N., Côme, E., Oukhellou, L., and Govaert, G. (2014). Clustering the vélib' dynamic origin/destination flows using a family of poisson mixture models. *Neurocomputing*, 141 :124–138.

Rijal, H., Tuohy, P., Humphreys, M., Nicol, J., Samuel, A., and Clarke, J. (2007). Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings. *Energy and Buildings*, 39(7) :823–836. Comfort and Energy Use in Buildings - Getting Them Right.

Samé, A., Chamroukhi, F., and Govaert, G. (2009). Modèle à processus latent et algorithme em pour la régression non linéaire. *41e Journée de Statistique, SFdS*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.

Stephen, B., Mutanen, A., Galloway, S., Burt, G., and Järventausta, P. (2014). Enhanced load profiling for residential network customers. *Power Delivery, IEEE Transactions on*, 29 :88–96.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3) :586–600.

Vázquez, F. I. and Kastner, W. (2011). Clustering methods for occupancy prediction in smart home control. In *2011 IEEE International Symposium on Industrial Electronics*, pages 1321–1328.

Wang, S., Chaganty, A. T., and Liang, P. S. (2015). Estimating mixture models via mixtures of polynomials. *Advances in Neural Information Processing Systems*, 28 :487–495.

Wasserman, L. (2010). *All of Statistics : A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., and Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, 82 :1027–1047.

Widén, J. and Wäckelgård, E. (2010). A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy*, 87(6) :1880–1892.

Wolf, S., Kloppenborg Mø ller, J., Bitsch, M. A., Krogstie, J., and Madsen, H. (2019). A markov-switching model for building occupant activity estimation. *Energy & Buildings*, 183 :672–683.

Xia, Y.-M. and Tang, N.-S. (2019). Bayesian analysis for mixture of latent variable hidden markov models with multivariate longitudinal data. *Computational Statistics & Data Analysis*, 132 :190–211. Special Issue on Biostatistics.

Yan, D., O'Brien, W., Hong, T., Feng, X., Burak Gunay, H., Tahmasebi, F., and Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation : Current state and future challenges. *Energy and Buildings*, 107 :264–278.

Yilmaz, S., Chambers, J., and Patel, M. (2019). Comparison of clustering approaches for domestic electricity load profile characterisation - implications for demand side management. *Energy*, 180.