Habitats, habitants et pratiques énergétiques : intégrer les situations d'habitation dans la modélisation quantitative de la consommation d'énergie domestique

Matthias Heinrich

THÈSE DE DOCTORAT

de l'École des Ponts ParisTech

Habitats, habitants et pratiques énergétiques : intégrer les situations d'habitation dans la modélisation quantitative de la consommation d'énergie domestique

École doctorale Villes, Transports et Territoires (VTT)

Doctorat en Génie Urbain

Thèse préparée au sein du laboratoire Techniques, Territoires, Sociétés (LATTS)

Thèse soutenue le 16/01/2024

Matthias HEINRICH

Composition du jury :

Marjorie Musy Directrice de Recherche, CEREMA	*Rapporteur*
Fateh Belaïd Chercheur, KAPSARC	*Rapporteur*
Faïcel Chamroukhi Professeur des Universités, Université de Caen	*Président*
Béatrice Roussillon Maitresse de conférences, Université de Grenoble-Alpes	*Examinatrice*
Jean-Pierre Lévy Directeur de Recherche, CNRS	*Directeur de thèse*
Latifa Oukhellou Directrice de Recherche, Université Gustave Eiffel	*Co-directrice de thèse*

Résumé

Cette thèse adopte une démarche interdisciplinaire pour produire une modélisation quantitative des consommations énergétiques basée sur les « situations d'habitation » (les relations ménages-logements), les pratiques domestiques et les contextes résidentiels. Elle est composée de trois parties.

Dans un premier temps, une revue de la littérature portant sur les principaux modèles des sciences de l'ingénieur et des sciences humaines permet d'identifier les cadres théoriques d'une approche associant les ménages et les logements (« situation d'habitation »). Dans le deuxième chapitre, les méthodes de modélisation des pratiques domestiques sont discutées, afin de déterminer les plus pertinentes pour étudier les liens entre les situations d'habitation et les styles de vies résidentiels. Dans le dernier et troisième chapitre, en s'appuyant sur un modèle de mélange de régressions à proportions logistiques, une modélisation hiérarchique de la consommation d'énergie domestique basée sur les situations d'habitation est proposée.

Située entre les approches typologiques et les approches de régression, l'un des apports de cette thèse est de proposer, dans une même méthode, une typologie de situations d'habitation et un modèle de régression. Les performances d'estimation des résultats sont équivalentes à celles des principaux modèles mais, en revanche, ils ont une portée explicative supérieure en associant les pratiques domestiques aux consommations d'énergie. Cette méthode permet également d'intégrer une segmentation des pratiques domestiques dans un modèle des consommations d'énergie à l'échelle du logement. La thèse comporte aussi une dimension heuristique. Elle montre tout d'abord les liens entre les situations d'habitation et les pratiques liées à l'équipement, l'occupation et les gestes de régulation. De plus, elle présente des situations d'habitation idéales-typiques mises en évidence à partir des caractéristiques des ménages, des logements, des comportements et les consommations d'énergie domestique totale, par personne et par mètre carré. Enfin, en termes prospectifs, elle ouvre la voie à la production d'autres approches interdisciplinaires portant sur la transformation des situations d'habitation.

Abstract

This thesis adopts an interdisciplinary approach to produce a quantitative model of energy consumption based on 'housing situations' (household-dwelling relationships), domestic practices and residential contexts. It consists of three parts. Firstly, a review of the literature on the main models from the engineering sciences and the humanities is directed to build a theoretical frameworks articulating households and dwellings ('housing situations'). In the second chapter, the modelling of the domestic practices is discussed, to determine which algorithm is the most relevant for studying the links between housing situations and residential lifestyles. In the final and third chapter, using a mixture model of regressions weighted by logistic proportion, a hierarchical modelling of domestic energy consumption based on living situations is proposed.

This research is situated between typological and regression approaches, and one of the contributions of this thesis is to propose, in the same method, a typology of housing situations and a regression model. The estimation performance of the results is equivalent to that of the main models but, on the other hand, they have a greater explanatory power by associating domestic practices with energy consumption. This method also makes it possible to integrate a segmentation of domestic practices into a model of energy consumption at the dwelling level. The thesis also has a heuristic dimension. Firstly, it shows the links between housing situations and practices relating to equipment, occupancy, and regulation. In addition, it presents ideal-typical living situations based on the characteristics of households, dwellings, behaviour, and total domestic energy consumption per person and per square metre. Finally, in forward-looking terms, it paves the way to produce other interdisciplinary approaches to the transformation of housing situations.

Avertissement

L’Ecole n’entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Remerciements

Au moment de l'écriture de ces remerciements, je repense à une conviction que je me suis construite durant ce travail de thèse qui est que le travail de recherche est mené souvent seul mais reflète d'abord les rencontres et les échanges qui l'ont jalonné. Ainsi, j'ai une pensée chaleureuse pour toutes celles et ceux, ami.e.s et camarades doctorant.e.s, enseignant.e.s, chercheuses et chercheurs qui ont entouré ce travail et m'ont accompagné durant mes réflexions et jusqu'au rendu de ce manuscrit. Je suis convaincu que les lignes qui suivront auront un peu des questions, des réponses, des exemples, des idées de chacun et de chacune d'entre elles : qu'ils et elles soient remerciés infiniment pour leur présence et leur soutien.

Je tiens tout d'abord à remercier Jean-Pierre Lévy et Latifa Oukhellou d'avoir accepté de diriger ma thèse. L'idée de faire collaborer sciences sociales et modélisation statistique sur la thématique des consommations d'énergie était ambitieuse et a généré de nombreuses problématiques pour nous comprendre, faire converger les questionnements et les méthodes mais je crois que nous aurons réussi à proposer une contribution intéressante. Merci pour votre confiance et votre soutien tout au long de ce travail.

Je joins à ces remerciements Marie Ruellan et Allou Samé qui ont pris une part très importante dans l'encadrement de ma thèse. La thèse a été menée à l'interface entre plusieurs disciplines (sciences de l'ingénieur, sciences de la donnée, et sciences humaines) et leur présence a été d'un grande aide pour mener ce travail à bien. Merci à Marie à qui je dois beaucoup. Merci pour son soutien ferme, sa présence chaleureuse et persévérante, son observation patiente et sa relecture qui m'ont aidé à tenir bon dans les orages du travail de thèse. Merci également à Allou Samé pour son encadrement rigoureux, son soutien, sa lecture et sa critique toujours bienveillante de mes propositions de modélisation parfois farfelues. La collaboration interdisciplinaire, bien que difficile d'emblée, et bousculée par la crise sanitaire a permis de proposer ce manuscrit qui témoigne des questions, des éléments de réponse et des éléments méthodologiques que nous avons apportés.

Je tiens également à remercier tous les membres du jury qui ont accepté de lire mon travail et d'en évaluer le contenu : je sais que cela représente un investissement en temps conséquent dans des agendas souvent déjà bien remplis. Aussi, la participation à un jury de thèse interdisciplinaire représente un défi supplémentaire et les critiques et les échanges qui ont suivi m'ont beaucoup aidé. Ils nourriront les publications qui suivront ce travail de thèse.

Ces remerciements seraient bien sûr incomplets sans mentionner l'équipe administrative du LATTS que j'ai eu le plaisir de côtoyer et de connaitre pendant ces 3 années. Merci à Nathalie, Assetou, Valérie, Nita, Noro, Aurélie, Virginie et Delphine pour votre aide et ces échanges qui ont égayé ces années passées au laboratoire. Merci aussi à l'équipe du GRETTIA, Marie-Laure et Mustapha pour leur soutien.

Je souhaite aussi remercier chaleureusement les jeunes et moins jeunes chercheuses et chercheurs du LATTS, du GRETTIA et du SATIE qui m'ont ouvert leur porte et avec qui j'ai apprécié échanger.

Merci aussi à l'équipe des coureurs de fond *lattsienne*, j'ai nommé Victor, Elise, Clarence, Emmanuelle pour ces heures de défoulement, merci aussi à Mariama, François, Olga, Mariana, Sofia, Paola, Roberta, Rina, Guillaume, Lauren, Youenn pour les moments partagés. Un très grand merci à Alexis et Marina pour nos échanges et les moments partagés. Je garderai un souvenir très chaleureux de chacun de vous et vous souhaite à chacun et chacune la meilleure des suites.

Merci aussi à mes amis pour leur soutien et ces précieux moments qui m'ont permis de sortir la tête de la thèse et de profiter de la Vie. Merci à Quentin, Antoine, Clémence, Matthias, Sébastien, Edgar, Baptiste, Florien, Guillaume. Avec chacun des qualités qui vous vont à merveille vous me donnez à chacune de nos retrouvailles un peu plus de joie.

Aussi, je souhaite remercier les personnes et instances qui ont contribué à faire que cette thèse ait pu advenir. Hamid et Bernard pour m'avoir soutenu et encouragé dès le premier jour de mon arrivée à l'ENS Rennes : je souhaite à chacun de pouvoir rencontrer des personnes aussi brillantes, bienveillantes et soucieuses d'apporter leur juste part pour accompagner intellectuellement et humainement leurs élèves. Je remercie aussi le corps des Ingénieurs des Ponts des Eaux et des Forêts qui a choisi de financer ce travail de thèse interdisciplinaire.

Enfin, je veux aussi remercier ma famille pour sa présence et sa confiance dans mon travail et ma manière de « mener ma barque ». La thèse n'a pas manqué de me poser de nombreuses questions. Votre patience et votre amour m'ont permis de franchir cette étape, d'apprendre beaucoup - sûrement, et d'apporter des éléments qui permettront des travaux et des discussions inter-disciplinaires - je l'espère !

Sigles

ACM	Analyse en Composantes Multiples
ACP	Analyse en Composantes Principales
ADEME	Agence de l'Environnement et de la Maitrise de la Demande d'Energie
AFDM	Analyse Factorielle de Données Mixtes
AIC	Akaike Information Criterion (Critère d'information d'Akaike)
BIC	Bayesian Information Criterion (Critère d'information bayésien)
CAH	Classification Ascendante Hiérarchique
DPE	Diagnostic de Performance Energétique
ECS	Eau Chaude Sanitaire
EPG/EEG	Energy Performance Gap/Energy Efficiency Gap
FEC	Final Energy Consumption
FECM2	Final Energy Consumption per square meter
FECP	Final Energy Consumption per household member
GES	Gaz à effet de serre
ICL	Integrated Completed Likelihood
MTE	Ministère de la Transition Ecologique
PR	Personne de Référence
SH	Situation d'Habitation
VIF	Variance Inflation Factor (Facteur d'inflation de la variance)
VS	Synthetic Variable (variable synthétique issue d'une classification de variables)
XGBoost	eXtrem Gradient Boosting

Sommaire

Introduction générale. « Rénover les logements » ou « transformer les modes d'habiter » ?

Le secteur résidentiel à l'épreuve des crises énergétique et climatique

A l'heure de l'écriture de ce manuscrit de thèse, le 6e rapport du Groupe intergouvernemental d'experts sur l'évolution du climat (GIEC) a été publié et rappelle aux gouvernements et aux populations les dynamiques du changement climatique et l'urgence à mettre en œuvre des solutions pour diminuer son impact et s'y adapter. Aussi, le contexte de post-crise sanitaire du COVID-19 marque le retour d'une inflation forte notamment sur l'énergie (inflation de +63% des prix du pétrole entre mi 2021 et mi 2022 d'après les chiffres de l'Institut National de la Statistique et des Etudes Economiques - INSEE[1]), ceci accompagnant une intensification de la guerre sur le continent européen et une insécurité croissante des chaines d'approvisionnement. Dans ce contexte, les gouvernements cherchent à réduire leur dépendance aux énergies fossiles par la transition vers les énergies renouvelables et décarbonées et à diminuer les consommations d'énergie. Le secteur résidentiel représente un gisement majeur d'économies. En effet, à l'échelle de la France la consommation d'énergie dans les logements a atteint en 2019 (hors tertiaire) près de 30% du total de l'énergie finale consommée en France (MTES 2022a) soit environ 487 TWh d'énergie finale[2]. Associée à cela, la combustion des énergies fossiles (fioul et gaz essentiellement) génère des émissions qui ont contribué à hauteur de 75 $MtCO_{2,eq}$[3], soit environ 17% des émissions de gaz à effet de serre en France la même année, ce qui en fait le 4e secteur le plus émissif derrière le transport, l'industrie et l'agriculture. Au niveau européen, les usages domestiques de l'énergie ont représenté en 2018 une consommation de 3267 TWh d'énergie finale (25% de l'ensemble), et 447 $MtCO_{2,eq}$ d'émissions de gaz à effet de serre (11% du total) (MTES 2022b). En ayant bien en tête la grande diversité des habitats et des différents modes de consommation dans le monde, on peut aussi observer les ordres de grandeur de ces indicateurs à l'échelle mondiale sur les consommations annuelles en énergie finale (15000 TWh hors bois[4], soit 13% du total) et les émissions associées (6,9 $GtCO_{2,eq}$ soit 16,6% de l'ensemble). Ces ordres de grandeurs d'émissions peuvent être mis en regard avec le budget carbone estimé par le GIEC, qui a indiqué que l'humanité avait en 2018 le droit d'émettre encore 420

---

[1]

L'énergie finale correspond à « l'énergie livrée au consommateur pour sa consommation finale (essence à la pompe, électricité au foyer, etc.) ». Une autre manière de compter la consommation énergétique consiste à passer par l'énergie primaire qui désigne quant à elle « l'ensemble des produits énergétiques non transformés, exploités directement ou importés. Ce sont principalement le pétrole brut, les schistes bitumineux, le gaz naturel, les combustibles minéraux solides, la biomasse, le rayonnement solaire, l'énergie hydraulique, l'énergie du vent, la géothermie et l'énergie tirée de la fission de l'uranium. ». (INSEE).

[3] Le $CO_{2,eq}$ correspond à une unité de mesure permettant de comparer les effets de différents gaz à effet de serre sur la base de leur potentiel de réchauffement global.

[4] Calculs personnels d'après le rapport de l'Agence internationale de l'Energie, disponible ici : https://iea.blob.core.windows.net/assets/1b7781df-5c93-492a-acd6-01fc90388b0f/Key_World_Energy_Statistics_2020.pdf

$GtCO_{2,eq}$ jusqu'en 2050 pour avoir 66% de chances de limiter le réchauffement global à 1,5°C en 2050, alors que les émissions annuelles mondiales sont d'environ 42 ± 3 $GtCO_{2,eq}$.

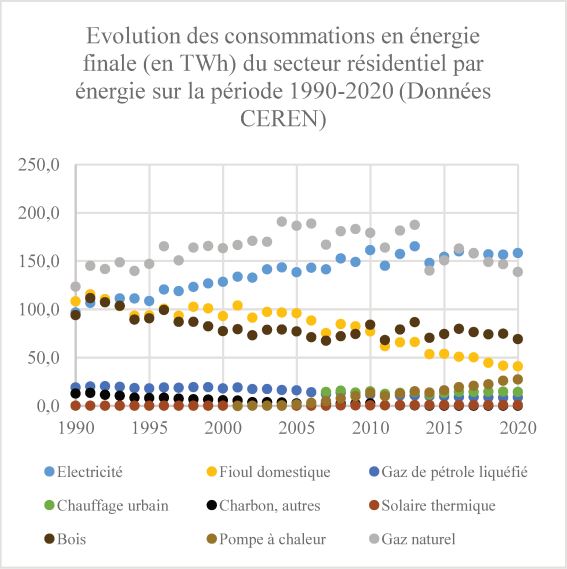
En conséquence, la France a pris des objectifs climatiques à travers l'Accord de Paris qu'elle a traduit dans la loi avec le vote de la Loi de Transition Energétique et pour la Croissance Verte (LTECV) en 2015. Celle-ci impose des objectifs chiffrés sur les émissions de gaz à effet de serre et les consommations en énergie finale et primaire, par exemple de « réduire les émissions de gaz à effet de serre de 40% entre 1990 et 2030 et diviser par quatre les émissions de gaz à effet de serre entre 1990 et 2050 ; réduire la consommation énergétique finale de 50% en 2050 par rapport à la référence 2012 en visant un objectif intermédiaire de 20% en 2030 ; porter la part des énergies renouvelables à 23 % de la consommation finale brute d'énergie en 2020 et à 32 % de la consommation finale brute d'énergie en 2030 ». La LTECV prévoit par ailleurs l'élaboration de deux outils : la Stratégie Nationale Bas Carbone (SNBC) et la Programmation Pluriannuelle de l'Energie (PPE). La SNBC est un document administratif élaboré par les services de l'Etat qui permet de donner des objectifs sectoriels de réduction des émissions de GES. Révisée en 2018, elle prévoit une réduction de -49% des émissions du secteur résidentiel en 2030 par rapport à 2015 et la neutralité carbone en 2050. Les moyens techniques présentés pour y parvenir sont la rénovation et l'isolation thermique massive des logements, le remplacement des énergies fossiles (fioul domestique, gaz) par des énergies décarbonées (pompe à chaleur, solaire électrique, nucléaire, solaire thermique, éolien, réseaux de chaleur, géothermie), et la promotion des pratiques plus vertueuses dans la construction et l'incitation au changement de comportement des habitants[5].

La rénovation thermique des logements comme moyen structurant des politiques climatique et énergétique

Lorsque nous nous intéressons aux usages domestiques en France en termes de volume d'énergie, nous observons que le chauffage occupe une place prépondérante (en moyenne entre 60 et 70% de l'énergie finale en 2019 d'après les données du CEREN[6] - voir Figure 1), suivi par la consommation d'eau chaude sanitaire (entre 10 et 15%), puis l'usage d'appareils électriques (environ 18%). Les énergies utilisées dans l'espace domestique sont par volumes décroissants l'électricité, le gaz, le bois, le fioul domestique puis en proportions moindres la pompe à chaleur, le chauffage urbain, le gaz de pétrole liquéfié, le solaire thermique et le charbon.

[5] La page d'informations du gouvernement sur la SNBC est accessible à ce lien : https://www.ecologie.gouv.fr/strategie-nationale-bas-carbone-snbc

[6] Le Ceren (Centre d'études et de recherches économiques sur l'énergie) est un groupement d'intérêt économique rassemblant des industriels de l'énergie (GRDF, ENEDIS, RTE) et des acteurs publics (ADEME, INSEE, Service des statistiques du Ministère de la Transition Ecologique – MTE).



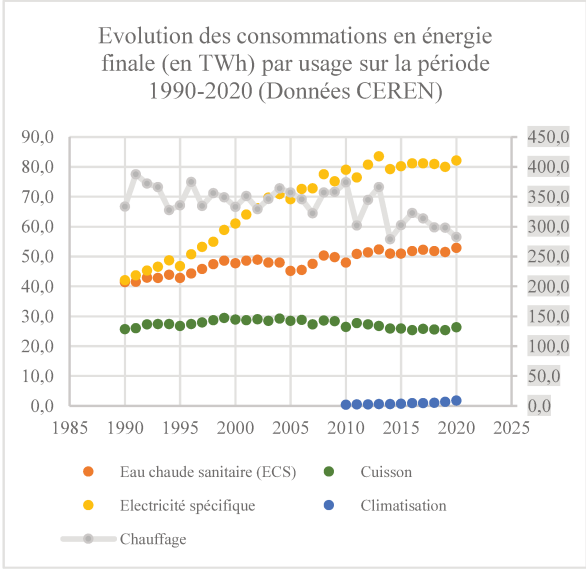


Figure 1 : Consommation en énergie finale par énergie (à gauche) et par usage (à droite). Les données en énergie finale sont issues de calcul du CEREN (voir note de bas de page). Notes de lecture. Graphe de gauche : les énergies finales sont données en PCI (Pouvoir calorifique inférieur). Graphe de droite : les données de consommation de climatisation n'ont été collectées que à partir de 2010. La consommation pour le chauffage est à lire sur l'axe secondaire situé à droite du graphique.

Cette observation est fondamentale pour comprendre l'orientation des politiques dans le bâtiment qui s'est structurée en trois temps depuis la première crise pétrolière de 1973. A cette date, la rareté du pétrole génère une chasse aux économies d'énergie dans les secteurs des transports, de l'industrie et du bâtiment. La première règlementation thermique du bâtiment est publiée en 1974 et établit de nouvelles normes de construction et d'isolation thermique, ainsi que de nouvelles normes de chauffage et de ventilation. La consommation de chauffage des logements est particulièrement ciblée : estimée jusqu'alors à plus de 300 $kWh_{ep}/m^2/an$ elle devait atteindre 225 $kWh_{ep}/m^2/an$ dans les nouveaux logements. En 1982, 1988 et en 2000 des mises à jour de la réglementation thermique relèvent les exigences de performance thermique (130 $kWh_{ep}/m^2/an$ dans la RT2000). Avec la RT 2005 puis la RT 2012 sont intégrés les calculs des consommations conventionnelles des plusieurs postes : le chauffage, de refroidissement, l'éclairage, la production d'eau chaude sanitaire et l'utilisation d'appareils auxiliaires (pompes et ventilateurs). Le plafond de consommation dans le neuf est alors de 50 $kWh/m^2/an$ en énergie primaire pour l'ensemble de ces consommations conventionnelles. Ces règlementations sont les premières à diriger les professionnels du bâtiment vers l'usage des énergies renouvelables. Enfin elles ont introduit de nombreux labels de construction dont HPE (Haute performance énergétique), THPE (Très haute performance énergétique), BBC (Bâtiment basse consommation) etc. Sur le plan financier, les années 2000 correspondent également aux premiers dispositifs mis en place par l'Etat pour encourager les rénovations des logements (l'éco-prêt à taux zéro dit « éco-PTZ », la TVA à taux réduit à 5,5% pour les travaux de rénovation énergétique, le Crédit d'impôt à la transition énergétique dit CITE). La dernière règlementation thermique dite RE2020 acte l'évaluation des bâtiments neufs sur des performances climatiques, des performances énergétiques sur le cycle de vie du bâtiment, et de confort

notamment en période de canicule. Avec la loi ELAN de 2018, le diagnostic de performance énergétique (DPE) prend un caractère juridique contraignant alors qu'il était utilisé jusque-là comme un outil informatif servant à évaluer la performance énergétique d'un logement. Enfin, la loi portant sur la lutte contre le dérèglement climatique et le renforcement de la résilience face à ses effets (2021) complète le dispositif de l'Etat en forçant le gel des loyers des logements non rénovés, et en interdisant progressivement la location des logements énergivores, aussi appelés « passoires énergétiques ».

La rénovation et le renouvellement du parc de logements français en retard

Ces mesures ont permis le lancement d'un marché de la rénovation thermique des bâtiments en stimulant la demande en travaux de rénovation. Le rythme de la rénovation, mesuré en nombre de logements rénovés par an est cependant largement insuffisant pour répondre aux objectifs chiffrés de la PPE et de la SNBC. D'après les données du ministère de la transition écologique, le parc résidentiel français se composait en 2021 de 37,2 millions de logements ordinaires (dont 1 million hors métropole), au sein desquels on différencie les résidences principales, les résidences secondaires et les logements vacants (respectivement 82%, 10% et 8% de l'ensemble). Le taux de croissance et le taux de renouvellement annuels sont respectivement de 0,8% et 1%. L'objectif de réaliser 500 000 « rénovations thermiques » (soit près de 1,3% du parc) chaque année a été communiqué par les autorités (notamment dans le Plan rénovation énergétique des bâtiments) mais le rythme réel apparaît bien en deçà pour de plusieurs observateurs. L'Agence Nationale de l'Habitat (ANAH) a indiqué en 2021 avoir financé 645 000 demandes d'aide mais a reconnu que 68% d'entre elles concernaient des travaux de changement de système de chauffage, et seules 21% concernaient l'isolation thermique du bâti. Le Haut Conseil pour le Climat pointait en 2020 dans son rapport « Rénover mieux : leçons d'Europe » (HCC 2020) que la France voit la performance énergétique de son parc progresser à une vitesse équivalente à celle des pays européens, mais insuffisante pour atteindre les objectifs qu'elle s'est fixée dans la SNBC. Le rythme des « rénovations complètes »[7] doit ainsi passer de 0,2% du parc en 2018 à 1% après 2022 et 1,9% en 2030.

La sobriété au secours des objectifs climatiques et énergétiques ?

Il est intéressant de noter qu'en parallèle du volet technique présenté, les discours publics ont aussi visé les individus et leur consommation d'énergie. Aussi, dès la crise pétrolière de 1973, le premier ministre Pierre Messmer fait appel à la « citoyenneté » pour économiser l'énergie. Depuis lors, les organismes publics ont diffusé des informations au public sur les impacts associés à chacun de leurs gestes et en particulier dans l'espace domestique. L'Agence pour les économies d'énergie (devenue Agence pour la Maitrise de la Demande Energétique – ADEME) qui fut créé à cette époque en est un relais important

[7] L'expression sert à distinguer les chantiers pour lesquels plusieurs travaux sont envisagés : isolation thermique du bâti, changement du système de chauffage et des fenêtres etc.

en France. On remarquera que les scénarios de trajectoire énergétique publiés par RTE[8], NegaWatt[9] ou l'ADEME[10] proposent des trajectoires énergétiques pour le pays, compatibles avec les engagements climatiques de l'Accord de Paris mais qui diffèrent dans l'importance relative de l' « efficacité énergétique » (la diminution des consommations d'énergie à service constant) et de la « sobriété énergétique » (diminution des consommations d'énergie par le changement de comportement). Ces *scenarii* présentent des visions divergentes de l'importance que doivent prendre les changements de comportement. La reprise du conflit en Ukraine au début de l'année 2022 a provoqué une remobilisation des pouvoirs publics, puisque l'Etat Français a publié un Plan de sobriété énergétique qui pour objectif une diminution de 10% de la consommation d'énergie finale en 2 ans et pour moyen dans le secteur résidentiel la réduction du chauffage à 19°C, le décalage de la fin et du début de la période de chauffage, la réduction de l'usage de l'eau chaude sanitaire (Gouvernement 2022). L'opérationnalisation des changements de comportements reste toutefois difficile et renvoie d'abord à une compréhension fine des logiques qui les sous-tendent.

L'articulation des comportements, des ménages et des logements au sein de modes de vies résidentiels

Les politiques énergétiques mises en place ont eu pour objectif de transformer le bâti, changer les équipements et les pratiques énergétiques des ménages. Les projets de rénovation thermique, le changement des comportements et des équipements interviennent cependant dans des cadres de vie préexistants en ayant une logique plus complexe que la seule consommation d'énergie. Par exemple, les pratiques de consommation domestiques s'intègrent dans un systèmes d'opportunités, de contraintes et de besoins déterminés notamment par les modèles culturels, les moyens financiers du ménage, l'environnement matériel du logement et des systèmes, les connaissances et les compétences et la structure du ménage. La prise en compte de ces éléments au sein d'une approche « habitante » transforme ainsi le regard et les « solutions » que l'on peut apporter pour diminuer les consommations d'énergie. Cette perspective qualifiée ici « d 'habitante » n'est toutefois pas le seul modèle permettant de comprendre l'origine des comportements et des consommations. Les recherches en sociologie, en économie, en psychologie ont fourni plusieurs théories fournissant à la fois une compréhension et des outils variés de la CED.

La diversité des approches explicatives de la CED

Les comportements énergétiques sont observables dans le quotidien des ménages : utiliser son four, laver son linge et le sécher suspendu sur un fil ou au sèche-linge, régler le thermostat sur 19°C, ouvrir une fenêtre (…). Toutefois, derrière l'évidence de l'observation se pose la question d'une possible

[8] Futurs énergétiques 2050 : les scénarios de mix de production à l'étude permettant d'atteindre la neutralité carbone à l'horizon 2050
[9] Scénario Négawatt
[10] «Transition(s) 2050. Choisir maintenant. Agir pour le climat »

interprétation : chacun de ces gestes peuvent-ils être compris isolément ou doivent-ils être intégrés dans un ensemble plus large (auquel cas quel serait son périmètre) ? Aussi, ces comportements sont-ils liés à un calcul rationnel des habitants ou plutôt d'une routine ? Dans quelle mesure ces comportements sont-ils conditionnés par le milieu social, les moyens financiers du ménage et les prix de l'énergie etc. ? En parallèle des recherches très nombreuses en économie et en psychologie, celles en anthropologie et en sociologie de l'énergie ont tenté d'étudier les manières d'habiter les logements en recherchant les logiques sous-jacentes. Pour ces travaux, l'hypothèse est que ces comportements pouvaient (et devaient) aussi être étudiés en lien avec les autres comportements, les caractéristiques des personnes, des ménages, du logement où ils étaient réalisés. Dans cette perspective, les usages intégrés dans la structure comportementale, sociale (les individus qui composent le ménage, mais aussi les groupes d'appartenance – la famille, les amis, la catégorie socio-professionnelle) et technique (les systèmes comme la télévision, les lampes, le four – l'environnement du logement) permettent d'expliciter une logique qui dépasse le simple geste observé. Cette perspective est dite *compréhensive* : c'est-à-dire qu'elle associe l'ensemble de ces éléments dans un discours explicatif cohérent. Il s'agit d'un glissement : le comportement « n'appartient » plus seulement à l'individu qui le réalise et au contexte où il est observé, il devient la réalisation ce qui est appelé une pratique sociale dans un contexte singulier. Le renouvellement de la recherche sur l'énergie dans l'espace domestique ainsi proposée a permis des apports théoriques et pratiques importants : identifier de nouveaux objets à étudier (des ensembles de comportements, des matériels, des discours, des parcours résidentiels), de nouveaux concepts (modes de vies, pratiques sociales) et des résultats importants pour décrire, comprendre et prévoir le changement des consommations d'énergie dans le logement. Sur le plan pratique, les travaux réalisés ont permis d'identifier des modèles culturels contextualisés qui permettent de mieux comprendre les logiques qui sous-tendent les comportements observés, à l'échelle de micro-études sur quelques dizaines de logements (comme dans Bonnin 2016 ; Subrémon 2009) ou sur plus larges (Garabuau-Moussaoui 2009). Cette formulation holistique des comportements trouve par ailleurs un écho dans les travaux récents qui cherchent à étudier les processus de transition (Eon et *al.* 2019; McMeekin and Southerton 2012 ; Shove, Walker, et Brown 2014).

Modéliser la transition des modes de vie résidentiels : quels sont les liens entre les situations d'habitation, les comportements et les consommations d'énergie ?

Cette approche reste toutefois relativement marginale dans la compréhension de la CED et dans la formulation des politiques publiques portant sur les comportements énergétiques dans le logement (Spurling et *al.* 2013). Une première raison est la faible acculturation dans de nombreux pays aux modèles sociologiques pour des raisons historiques et scientifiques (Delahais et Devaux-Spatarakis 2018). Une seconde raison est que ces approches disposent encore de peu de travaux à des échelles spatiales plus importantes que les cas d'études réalisés sur des dizaines voire des centaines de logements (Bonnin 2016 ; Subrémon 2009). La « généralisation » du modèle se heurte toutefois à une difficulté

majeure qui est la haute complexité du modèle, qui dépend par exemple de variables relatives aux individus, aux caractéristiques physiques et symboliques du logement, à la composition du ménage, etc. La généralisation supposerait que l'on dispose d'un modèle exhaustif, éventuellement sous une forme mathématique et qu'en collectant des données sur une échelle plus large il deviendrait possible de le valider empiriquement. On voit bien à ce niveau la double difficulté (indisponibilité de telles données, complexité du modèle mathématique associé) que pose cette approche. Il peut toutefois être intéressant de revoir nos exigences : ce que propose ce modèle sociologique étant d'observer l'interrelation des variables de comportements et de contexte, il peut être intéressant pour les acteurs publics et privés agissant à des échelles plus larges que celles du logement d'observer quels sont les effets produits par cette interrelation forte à l'échelle d'un ensemble de logements ?

La proposition de la thèse : explorer l'influence des contextes résidentiels sur la CED

Ainsi, dans ce travail de thèse nous souhaitons explorer une modélisation multidisciplinaire de la CED. Notre objectif est d'y intégrer à la fois les dimensions techniques (nombre et qualité des équipements et du bâti), et comportementales. Les comportements seront abordés comme une somme de gestes consommateurs d'énergie étant liés dans une logique définie et contrainte par le ménage, les individus qui le composent et le logement.

Finalement, ce travail de thèse propose d'adopter une perspective systémique issue de la sociologie de l'énergie pour comprendre les liens entre logements, ménages, comportements et la consommation énergétique domestique. Il tente de fournir des éléments de réponse aux deux questions suivantes :

En quoi une modélisation quantitative de la consommation d'énergie domestique basée sur l'étude des contextes résidentiels permet-elle de contribuer à comprendre les liens entre les situations d'habitation, les comportements des habitants et les consommations d'énergie ?

Dans le contexte d'une transition énergétique et climatique du secteur du logement français, quels éclairages cette approche de modélisation peut-elle apporter à la connaissance des consommations énergétiques domestiques ?

Pour présenter ce travail de modélisation de CED, nous proposons de revenir dans le chapitre 1 (sur l'ensemble des modèles existants de CED à l'échelle du logement. Dans le chapitre 2, nous reviendrons sur les travaux empiriques ayant décrit des logiques habitantes. Nous les utiliserons pour réaliser une modélisation à partir d'une classification de variables comportementales que nous mettrons en relation avec celles qui caractérisent le logement et le ménage. Dans le chapitre 2, en faisant l'hypothèse que l'interaction entre ménage et logement déterminent en grande partie les « systèmes d'habitation », nous réaliserons une classification des ménages et des logements que nous nommons « situations d'habitation ». Nous étudierons la sensibilité de ces résultats à l'algorithme utilisé et valoriserons cette

méthodologie pour décrire, dans le Chapitre 3, deux applications de ce concept pour la modélisation de la CED aux échelles du logement et du parc immobilier en France métropolitaine.

Chapitre 1. Construction d'un cadre de modélisation de la CED à partir des situations d'habitation

> *« Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful. »*
> *Box, G. E., & Draper, N. R. (1987).*

La modélisation de la consommation d'énergie domestique (CED) entreprise dans ce travail de thèse s'inscrit dans une littérature abondante et multidisciplinaire. Nous visons dans ce chapitre à rendre compte de la diversité des théories mobilisées, des modèles construits et des variables identifiées avant de formuler une proposition de contribution qui sera testée dans une approche globale et numérisée. Nous faisons le choix de nous intéresser ici aux modèles de CED à l'échelle spatiale du logement, entendu comme un « local physique à usage d'habitation séparé des éventuelles parties communes et indépendant » (INSEE), tandis que la CED est considérée comme la consommation d'énergie annuelle associée au logement, tous combustibles confondus (gaz, gaz liquide, électricité, bois, fioul domestique, réseaux de chaleurs, excepté les ressources renouvelables en raison de données non disponibles).

La modélisation est un terme polysémique qui renvoie généralement à une approche déductive, réductrice, et souvent quantitative. Dans les faits, elle est une méthode aux objectifs variés dépendants des disciplines qui l'utilisent (Varenne and Silberstein, 2013). C'est notamment le cas dans le domaine de la modélisation de la consommation d'énergie résidentielle où les approches modélisatrices relèvent de courants scientifiques issus de l'économie, de la psychologie, de l'ingénierie, des sciences de la donnée mais aussi de l'anthropologie, de la sociologie de l'énergie (Lévy et Belaïd, 2018a ; Swan et Ugursal, 2009). Certains modèles de sciences de l'ingénieur ont pour objectif de prédire la consommation de nouveaux logements (Ahmed Gassar, Yun, and Kim 2019 ; Amasyali and El-Gohary 2018) ; d'autres servent à identifier les paramètres déterminants de la CED (Hansen, Gram-Hanssen, et Knudsen, 2018 ; Andersen, 2012) ou ses processus (Lutzenhiser, 1992 ; Frederiks, Stenner, et Hobman 2015 ; Van Raaij et Verhallen, 1983) ; enfin, elle peut servir d'appui aux politiques publiques (Giraudet, Guivarch, et Quirion, 2012 ; Glotin et *al.,* 2019), par exemple dans l'aide à la décision de travaux de rénovation (Serrano-Jiménez et *al.* 2021), etc.

Ces modèles diffèrent également dans leur définition des usages énergétiques. Une première hypothèse serait de les considérer comme une approche physique ou mathématique visant à produire des archétypes standardisés des usages énergétiques[11] en reproduisant statistiquement ou mathématiquement une

[11] Voir par exemple l'Arrêté du 6 mai 2008 portant confirmation de l'approbation des diverses méthodes de calcul pour le diagnostic de performance énergétique en France métropolitaine – Journal officiel, 2008. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000018801075

diversité observée dans des enquêtes (Vorger 2014 ; Buttitta et *al.* 2019). Les approches qualitatives, quant à elles, s'appuient sur des méthodes plus inductives avec l'objectif de modéliser les processus de comportements énergétiques, en ignorant généralement l'influence relative de l'efficacité énergétique du bâti et des systèmes. Cette représentation binaire de la modélisation, entre quantitatif et qualitatif, est cependant trop simplificatrice, mais elle a le mérite de permettre d'introduire notre l'état de l'art de la question.

Cette dichotomie renvoie à des approches imperméables. C'est la raison pour laquelle notre thèse porte sur une tentative de résoudre le verrou d'une intégration des usages dans un modèle quantitatif des CED. A cette fin, nous nous proposons dans un premier temps de réaliser un inventaire des modèles existants et des propositions de modélisation multidisciplinaire les plus récentes ayant permis des avancées heuristiques et applicatives majeures dans l'étude de la CED.

1. Un inventaire multidisciplinaire des modèles de la consommation annuelle d'énergie domestique

1.1 Un rappel sur les modèles et la modélisation

Dans la question initiale de cette thèse, le terme de « modélisation » ne me semblait pas appeler à une définition. Quoi de plus évident pour un ingénieur que de construire un « modèle » ? Les discussions au sein de l'équipe de recherche, mais surtout les blocages dans les échanges multidisciplinaires, ont mis en avant la polysémie d'un terme à forte portée disciplinaire. En m'appuyant en grande partie sur les travaux de sociologie de la quantification d'Alain Desrosières et d'épistémologie de Franck Varenne, je propose de décrire d'abord une vision transversale des débats et des approches sur la modélisation pouvant être mobilisés dans le cadre de la CED.

Une définition du « modèle » est-elle possible ?

Ainsi que Franck Varenne le rappelle dans ses travaux (Varenne 2008), le terme de « modèle » vient du latin « modulus » qui peut être traduit par « petite mesure ». Un modèle est donc en quelque sorte un objet qui permet de mettre en rapport des éléments (comme un mètre étalon qui permet de comparer des éléments d'une architecture). Le terme de « modèle » peut ainsi faire référence tantôt « à l'objet imité, tantôt à l'objet qui imite » (Varenne 2008, p.3). Surtout, cette première définition met en avant le caractère relationnel du « modèle » ; entre un objet construit et un phénomène réel. L'enjeu est alors de décrire et comprendre les formes, les enjeux, les implications, les avantages et les risques que font peser la construction de ces objets. Le modèle est-il forcément un ensemble d'équations mathématiques ? Ou peut-il prendre des formes matérielles (une maquette n'est-elle pas aussi un modèle ?) ? Aussi peut-on faire des modèles sans théorie (et qu'est donc une théorie sinon une forme de modèle ?) Et comment peut-on comprendre le rôle des simulations qui se sont multipliées avec l'essor de l'informatique : ne sont-elles pas elles aussi des modèles qui permettent de produire des résultats quantitatifs en écartant la

résolution analytique des modèles mathématiques traditionnellement développés en physique classique ?

Dans son travail, Varenne illustre la diversité des formes et des objectifs remplis par les modèles, qui dépendent des contextes sociaux et historiques. Il illustre par ailleurs la diversification et la complexification observée des pratiques de modélisation (en forme, en fonctions). Ainsi, plutôt qu'une définition, l'auteur s'appuie sur la caractérisation minimale des modèles proposée par Minsky en 1965 (Minsky 1965).

> *Pour un observateur B, un objet A* est un modèle d'un objet A, dans la mesure où B peut utiliser A* pour répondre à des questions qui l'intéressent au sujet de A (Minsky, 1965 cité par Varenne, 2008 p.7)*

Une typologie des fonctions sociales des modèles

Cette caractérisation minimale permet à Varenne de tirer quatre enseignements communs aux « modèles » : il s'agit (1) d'objets (physiques ou non) autonomes de l'objet modélisé, au sens où ils existent en eux-mêmes. Ces objets (2) ne sont pas nécessairement construits par analogie ou par objectif de représentation, mais (3) leur définition est faite dans une double relativité avec l'objet modélisé et avec le modélisateur. Enfin, (4) ces objets sont construits avec l'objectif de répondre à question posée par le modélisateur.

Franck Varenne complète sa description en construisant une typologie des fonctions des modèles. Une synthèse de cette typologie en cinq fonctions est proposée dans le Tableau 1.

Tableau 1 : Synthèse des fonctions assumées par les modèles. Source : Travail de l'auteur à partir de données issues de (Varenne, 2008).

Fonction du modèle	Description
Faciliter l'accession au réel	Un modèle peut servir à faciliter une accession à ce qui se donne de manière sensible, mesurable ou détectable. Il peut ainsi faciliter une observation, une visualisation, mais aussi le rendu d'une expérience ou d'une expérimentation. **Exemple** : une maquette d'un pont
Rendre intelligible	Un modèle peut faciliter une présentation intelligible de l'objet modélisé via une représentation mentale ou une conceptualisation. **Exemple** : le calcul de statistiques agrégées (moyennes, écart-types)

Médiation objet-chercheur dans l'élaboration des connaissances	Le modèle sert de médiateur entre le chercheur ou une communauté de recherche et le réel, dans l'entreprise de compréhension voire de théorisation. **Exemple** : l'amélioration du modèle orbital de l'atome proposé de Rutherford, par Niels Bohr permet d'expliquer les spectres de raies de l'hydrogène (1913), et bien qu'il ait montré rapidement des limites, il a posé les bases de la mécanique quantique.
Médiation et circulation de la connaissance	Le modèle peut aussi servir à synthétiser et présenter un ensemble de connaissances disciplinaires pour les faire circuler d'un domaine disciplinaire à un autre. Le modèle sert de représentation simplifiée du réel et de sa représentation par une communauté en particulier : dans cette perspective, le modèle permet de véhiculer des représentations. **Exemple** : modèle multidisciplinaires du climat utilisé dans le rapport du GIEC. Chaque « brique » du modèle propose une représentation de l'évolution des systèmes liés au climat (atmosphère, océan, cryosphère, terre, biosphère, cycle du carbone).
Appui à la décision	Le modèle peut aussi être un instrument d'aide à la décision. En ciblant le ou les phénomènes en jeu, il n'est pas nécessairement précis voire exact car son objectif est d'abord d'être mobilisé comme support à la construction d'une politique. **Exemple** : modèle de gestion du risque financier ; modèles prédictifs de consommation d'énergie permettant à l'opérateur de gérer l'équilibre du réseau électrique entre l'offre et la demande.

Un élément particulièrement intéressant dans l'approche de Varenne est qu'elle dépasse la critique légitime et courante des modèles, qui en gagnant en complexité (souvent mathématique) s'écartent du réel. L'auteur souligne le fait que ce n'est pas tant les objets qui sont désignés par le mot « modèle », mais plutôt un certain type de « facilitation ». Clarifier la notion revient alors à illustrer la diversité des formes et surtout des objectifs des pratiques de modélisation. Cela permet de mieux comprendre l'apport ontologique, social, technique qu'apportent les objets modélisés dans leurs contextes. Cette vision peut alors être utilisée afin d'appréhender la diversité et la convergence des formes de modélisation de la CED dans les littératures en sciences humaines, en sciences de l'ingénieur et en sciences de la donnée.

Différences et rapprochements entre « modèle » et « théorie »

Dans ce cadre, il est utile de rappeler le rôle de la théorie et des simulations vis-à-vis de la pratique de la modélisation. Pour Varenne, la théorie désigne « un large ensemble d'énoncés – éventuellement formalisés et axiomatisés –, reliés systématiquement entre eux, formant système donc, et donnant lieu à des inférences susceptibles de valoir de manière générique pour un type de phénomènes donné » (Varenne, 2008 p. 11)[12]. Cette définition met en avant une différence fondamentale entre modèle et théorie car une théorie ne répond pas à une classe de fonctions[13]. Clarifier le statut du « modèle » permet d'observer le caractère profondément fonctionnel de ces objets et d'en appréhender la diversité des formes et des pratiques de construction.

Sur la problématique de la compréhension de la consommation d'énergie domestique qui dirige ce travail de thèse, nous avons observé une pluralité de questionnements, de théories et d'outils issus des différente disciplines. Dans la revue de littérature qui suit, il s'agit alors de rendre compte de la diversité des « modèles » (dont nous enlèverons à présent les guillemets) proposés par les différentes communautés scientifiques. Cela passera par l'explicitation des fonctions (appui à la décision, rendre intelligible, médiation, etc.), des formes (mathématiques, symboliques, etc.), et des théories mobilisées. Nous ordonnons la présentation des modèles en plusieurs parties. D'abord les modèles quantitatifs sont introduits en distinguant les modèles statistiques et les modèles basés sur des approches d'ingénierie. Ensuite les modèles qualitatifs issus de l'économie, de la psychologie et de la sociologie sont introduits. L'idée est pour le lecteur de mieux comprendre la place donnée à chacun des ensembles de variables (Figure 2). Les caractéristiques des logements, des ménages et des individus, les pratiques domestiques et les consommations d'énergie occupent une place plus ou moins importante dans les travaux de modélisation.

[12] A titre d'exemple, la théorie de l'acteur rationnel est utilisée dans le champ économique pour expliquer les comportements des agents. La théorie postule notamment le fait que les décisions d'agir sont fondées sur des inférences logiques à partir d'informations disponibles et de préférences.
[13] Attention, le terme de fonction renvoie ici à l'utilisation faite du modèle. Voir Tableau 1.

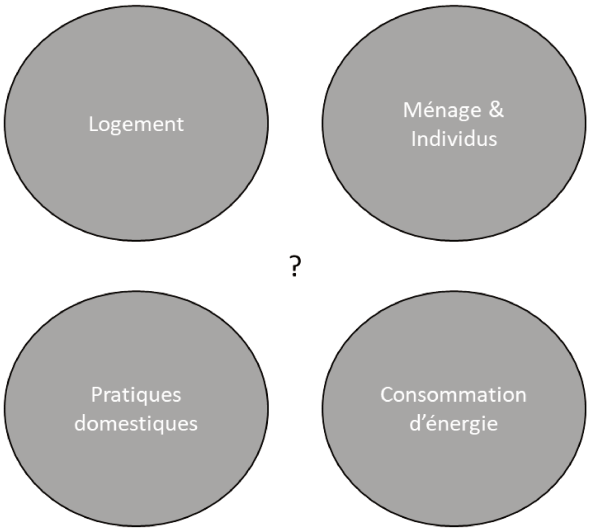


Figure 2 : Les caractéristiques des logements, des ménages et des individus, les pratiques domestiques et les consommations d'énergie occupent une place plus ou moins importante dans les travaux de modélisation. Source : Auteur.

1.2 Les modèles quantitatifs

1.2.1 Les modèles construits par apprentissage statistique

Une première grande famille de modèles regroupe les « modèles construits par apprentissage statistique ». Il s'agit de modèles mathématiques qui lient de manière linéaire ou non linéaire[14] un ensemble de variables explicatives à la variable expliquée, par exemple l'énergie finale consommée dans un logement ; l'énergie primaire ; l'énergie finale pour le chauffage ou pour un autre poste de consommation. Le modèle attribue à ces variables des coefficients dont la valeur est calculée par « entraînement », c'est-à-dire que leur valeur est déterminée par une suite d'opérations mathématiques qui visent à maximiser (ou minimiser) un critère de performance du modèle sur un ensemble de données. Par exemple, l'entraînement d'un modèle de régression linéaire vise à calculer la valeur des coefficients de régression. Celles-ci sont calculables par exemple en utilisant le critère dit des « moindres carrés » qui vise à minimiser l'écart entre les valeurs réelles et les valeurs calculées par le modèle. Dans cet exemple le calcul est très simple car on connaît une expression explicite (analytique) des coefficients mais ce n'est pas le cas en général où des estimations des valeurs sont calculées à l'aide de techniques mathématiques avancées. Ces modèles visent deux objectifs principaux : une visée explicative – c'est-à-dire l'identification de variables ayant un impact significatif sur la consommation d'énergie domestique (CED), ou une visée prédictive – c'est-à-dire l'estimation, avec une erreur raisonnable, de la consommation d'énergie dans une situation nouvelle. Il est important toutefois de rappeler que ces

[14] Un modèle dit linéaire est un modèle qui fait l'hypothèse que la variable expliquée (exogène), par exemple la consommation d'énergie d'un logement, est issue de l'addition d'effets supposés indépendants (endogènes), comme l'effet de la surface ou du nombre de personnes. Chacun des effets est supposé être proportionnel à la variable endogène. Un modèle non linéaire est un modèle qui ne respecte pas ces deux conditions

calculs ne sont pas forcément basés sur une transcription d'une compréhension d'un processus physique dans un langage mathématique, mais plutôt que ce sont des structures mathématiques construites à partir de données empiriques (enquêtes, mesures issues de capteurs, de factures etc.). Elles offrent la possibilité *a minima* d'identifier des corrélations entre consommations et caractéristiques du contexte, des logements et des ménages. Pour organiser la présentation des modèles nous différencions ceux qui sont linéaires et non linéaires, et ceux qui appréhendent la variable expliquée – l'énergie – comme une variable quantitative (on parle alors de régression) ou qualitative (classification). On différencie également les approches dites supervisées (le modèle est construit sur un ensemble de données où la variable expliquée est présente) et les approches non-supervisées (où la variable expliquée est absente).

Les modèles de régression linéaire

Trois stratégies de modélisation linéaires ont été développées : la régression multilinéaire, l'analyse conditionnelle de la demande et le modèle d'équations structurelles. Ces trois méthodes ont permis d'étudier les liens entre les variables et la consommation d'énergie dans les logements.

La régression multilinéaire

L'approche multilinéaire ou régression linéaire multiple est un modèle mathématique qui part de l'hypothèse qu'une variable expliquée (dite variable *endogène*) peut être estimée à partir d'un ensemble d'éléments explicatifs indépendants (dits aussi *exogènes*). Le modèle appliqué à la consommation d'énergie dans le logement s'appuie généralement sur une régression du logarithme de la CED (noté $log(CED)$ dans la suite du manuscrit). En considérant une notation stochastique où $Y = \log(CED)$ est la variable aléatoire à expliquer et $(X)_{j\in[[\,1,n]]}$ les variables aléatoires explicatives (pouvant être qualitatives ou quantitatives), on peut exprimer le modèle théorique suivant :

$$log\,(CED) = Y = a_0 + \sum_{j\in[[\,1,m]]} a_j\,X_j + \varepsilon$$

ε désigne le terme d'erreur qui modélise l'information non comprise dans les variables explicatives X_j. Les termes $(a)_{j\in[[\,1,n]]}$ désignent les paramètres constants, modélisant l'effet moyen de chacune des variables explicatives sur Y. Ces paramètres sont estimés à l'aide de la méthode des moindres carrés (Sergent et *al.* 1995). L'analyse est effectuée en trois temps. D'abord par une validation des hypothèses (homoscédasticité des résidus ε, i.e. indépendance entre les résidus et les variables explicatives, normalité des résidus). Dans un deuxième temps, par l'observation de la variance de l'estimation des paramètres (par exemple à l'aide du facteur d'inflation de la variance – acronyme VIF en anglais), et de la significativité statistique des effets identifiés (à travers la valeur de la probabilité de la non-nullité des coefficients aussi connue comme la « p-valeur »). En dernier lieu, par une estimation de la capacité prédictive du modèle en utilisant par exemple des indicateurs comme l'erreur quadratique moyenne sur de nouvelles données. La difficulté majeure de ce type d'approche se situe au niveau de la forte

corrélation entre les variables qui augmente l'incertitude sur les effets des indicateurs calculés et donc leur significativité statistique (Tsanas et Xifara, 2012). Toutefois, en dépit de cette limite, un grand nombre de modèles multilinéaires de la CED ont été proposés. Les études ont porté sur l'influence de variables décrivant le logement, le ménage, les comportements domestiques, les attitudes etc. Le Tableau 2 propose un inventaire non exhaustif des variables mobilisées dans la littérature pour construire des modèles linéaires de la CED. Des exemples d'ordre de grandeur des coefficients ou des effets en pourcentage des variables sont donnés dans le cas de la France.

Tableau 2: Inventaire non exhaustif des variables utilisées dans des modèles de régression linéaire multiple. Pour plus de clarté dans l'exposé, seuls les travaux ayant porté sur des étude de cas français sont listés. Les ordres de grandeurs sont donnés à titre indicatifs : leur variance est relativement importante selon l'espace géographique, la sélection des variables, de l'échantillon et la temporalité considérés.

Variables	Fréquence de la significativité dans la littérature (*** : Haute ; ** : moyenne ; * rare)	Exemples d'effets calculés (Ordres de grandeur)
Logement		
Type de logement (individuel/collectif)	***	Les logements collectifs ont une CED inférieure de -18% par rapport aux logements individuels, toute choses égales par ailleurs (MTES 2017)
Nombre de logements dans le bâtiment	***	-2% par logement supplémentaire dans le bâtiment (Risch et Salmon 2017)
Qualité de l'isolation thermique des murs	***	-14% en moyenne des logements ayant un DPE de A ou Bou C par rapport aux logements ayant un DPE de F ou G (MTES 2017)
Qualité de l'isolation thermique des fenêtres	*	-5% pour une isolation en double vitrage par rapport à un simple vitrage (Risch et Salmon 2017)
Surface (m²)	***	Elasticité de 0,38 pour les maisons individuelles et 0,55 dans les logements collectifs (MTES 2017)
Date de construction	***	Par rapport aux logements construits avant 1948, en moyenne : -4,5% (logements entre 1949 et 1989) et -20% (logements entre 1990 et 2005) (MTES 2017)
Type d'énergie utilisée pour le chauffage principal	***	En moyenne, par rapport à la catégorie chauffage « autre » : +32% (fioul), -30% (électricité), +15% (gaz) (MTES 2017)
Chauffage collectif	***	-40 % pour le chauffage collectif comparé à un autre type de chauffage, dans un logement collectif (MTES, 2017).
Climat (Degré Jour unifié ou autre)	***	Elasticité de 0,38 en moyenne (MTES 2017)
Exposition du logement (Nord/sud)	*	-4% sur la CED/m² pour une bonne exposition (Risch et Salmon 2017)
Ménage		
Nombre de personnes	***	Par rapport à un ménage composé d'une seule personne : +22% (2 personnes) et +34% pour 3 personnes et plus (MTES 2017)
Revenus du ménage	***	+ 10 % entre D1 et D10 (Cavailhès et *al.* 2011)
Âge de la personne de référence (PR)	**	+4%/10 ans – La différence est très significative pour les classes d'âge les plus élevées (+13% de différence entre la classe >65 ans et la classe <35 ans) (MTES 2017)
Catégorie socio-professionnelle de la PR	*	En prenant comme référence les professions intermédiaires, seules les catégories « agriculteurs », « artisans, commerçants,

		chefs d'entreprise » et « employés » sont significatives (resp. +22%, +12% et +7%).
Comportement		
Température de chauffe déclarée	***	+5%/°C (MTES 2017)
Nombre d'équipements domestiques	***	Elasticité de 0,5 (MTES 2017)
Période d'absence / occupation moyenne du logement	***	-10% pour les ménages s'absentant plus de 8h par jour par rapport aux ménages absents moins de 4h (MTES 2017)
Autre		
Prix de l'énergie	***	- 0,5 en maison, - 0,87 en appartement (Risch et Salmon 2017)
Aire urbaine	***	+15% (resp. -6%) pour les logements dans les aires urbaines de moins de 10 000 habitants (resp. >100 000 habitants) (MTES 2017)
Statut d'occupation du logement	***	-9% pour les propriétaires en moyenne (MTES 2017) mais les résultats sont variables ; (Risch et Salmon 2017) trouve le même chiffre sauf +15% pour les propriétaires occupants chauffés à l'électricité.

Le nombre important d'études portant sur cette structure mathématique sur la base de corpus de données très variés (dans le domaine, le nombre de variables et la taille des échantillons) nous fournit plusieurs enseignements. Elles confirment l'influence significative sur la CED des caractéristiques physiques du logement, de la composition du ménage, de ses revenus et de ses comportements. Ainsi, la surface et le type de logement, le mode d'énergie utilisé pour le chauffage principal, la composition du ménage, le revenu, l'âge de la personne de référence, la date de construction du logement et son isolation, le nombre d'équipements domestiques, le climat local apparaissent comme des déterminants dans la plupart des études. D'autres variables ont des effets contrastés selon la localisation, la taille de l'échantillon et sa représentativité : le statut d'occupation, la possession d'appareil de régulation, les attitudes exprimées sur la protection de l'environnement et la régulation des consommations, les catégories socio-professionnelles, le niveau de diplôme apparaissent comme des variables plus ou moins discriminantes, ce qui laisse ouvert le champ de recherche sur leur importance dans les comportements de consommation. En revanche, ces modèles restent souvent incomplets (la variance expliquée[15] par les modèles du logarithme de la consommation d'énergie se situe souvent entre 50 et 70%, et aux alentours de 50% s'agissant de la consommation non normalisée exprimée en kWh). L'ajout de variables explicatives diminue généralement en valeur absolue les effet moyens « toutes choses égales par ailleurs » des variables initiales (MTES 2017 ; Zhao et *al.* 2017). Aussi, la forte corrélation observée des variables explicatives peut être un frein à la robustesse de l'interprétation de leurs effets sur les comportements énergétiques. Enfin, la non-linéarité et les interactions ne sont pas traitées systématiquement dans cette approche bien qu'elles soient identifiées dans la littérature (ex : effet de l'âge non linéaire, interaction entre comportements et caractéristiques des ménages (Zhao et *al.* 2017 ;

[15] Pour rappel, la variance expliquée est un critère quantitatif entre 0 et 1, souvent exprimé en pourcentage et qui permet d'évaluer la qualité du lien statistique entre deux grandeurs quantitatives (par exemple X et Y). Il est calculé comme le carré du coefficient de corrélation de Brabais-Pearson entre les variables X et Y Si la variance expliquée de Y par X est proche de 100% (resp. 0%), la connaissance de la variable X permet (resp. ne permet pas) de connaître le score de Y avec une erreur raisonnable.

Henley et Peirson, 1998). Ainsi, des chercheurs ont construit des modèles sur la base de régressions sur des sous-ensembles homogènes de logements, de ménages, de comportements, de modes de vie ou sur des zones géographiques restreintes. Le choix des variables est effectué soit en raison de la qualité des données, soit pour analyser les comportements d'une population spécifique. Par exemple, (Lutzenhiser et Bender, 2010) ont réalisé une régression multiple sur cinq zones climatiques de la Californie du Nord et sur l'ensemble de la région. La comparaison des résultats montre que l'impact des jours de chauffe (Heating Degree Day – HDD) apparaît significatif dans une seule zone ; le revenu peut ne pas être discriminant dans une zone tout en l'étant dans d'autres. (Zhao et *al.,* 2017) ont, quant à eux, utilisé 14 variables (énergie consommée, équipements, comportements) issues d'un corpus de 312 ménages vivant dans des maisons multifamiliales de l'Etat de Virginie. Ils ont construit une méthode de calcul constituée d'une succession de modèles de régression, en ajoutant pas à pas des variables décrivant des équipements, des comportements domestiques, puis leur produit (afin d'estimer l'impact global de leurs interactions). Dans leurs conclusions, ils avancent qu'en moyenne les « avancées technologiques introduites dans l'espace domestique ne contribuent directement qu'à hauteur de 42% du potentiel » apporté par les « bâtiments verts » et qu'en conséquence la diminution des consommations d'énergie engageait à la fois des « avancées techniques et de la plasticité comportementale » (p.231).

Le modèle d'équations structurelles

Une approche alternative pour étudier les interactions entre les différentes variables est la modélisation par équation structurelle (ou *Structural Equation Modeling* – SEM). Cette modélisation, très utilisée en sciences sociales, nécessite de mobiliser des ressources théoriques pour formaliser un schéma de causalité impliquant des variables non observées (appelées *variables latentes*, Figure 3). L'ajustement du modèle par régression permet ensuite sa validation et son analyse. Dans le cas de la CED, des travaux comme ceux de (Belaïd, 2017 ; Estiri, 2015) mettent en évidence des exemples d'interactions. Dans la formulation des travaux de SEM, la CED est expliquée par des effets directs (observables) et des effets indirects (non observables). En particulier, ils font l'hypothèse que les caractéristiques des ménages déterminent en grande partie les caractéristiques des logements occupés et, qu'en conséquence, la part expliquée par les caractéristiques des logements dans la CED est surestimée (Figure 3). Selon ces auteurs, la construction de variables latentes, états-uniennes pour Estiri et françaises pour Belaïd, met en évidence le fait que si les caractéristiques du bâtiment (type, surface, type d'énergie) sont effectivement les premiers déterminants de la CED, les caractéristiques des ménages, croisées avec les choix résidentiels, ont un impact du même ordre de grandeur sur la CED. En revanche, dans ces modèles la variable latente modélisant les comportements énergétiques domestiques (modélisés par Belaïd à travers 10 variables décrivant l'occupation du logement et la régulation du chauffage) est significative mais a un effet mineur sur la CED.

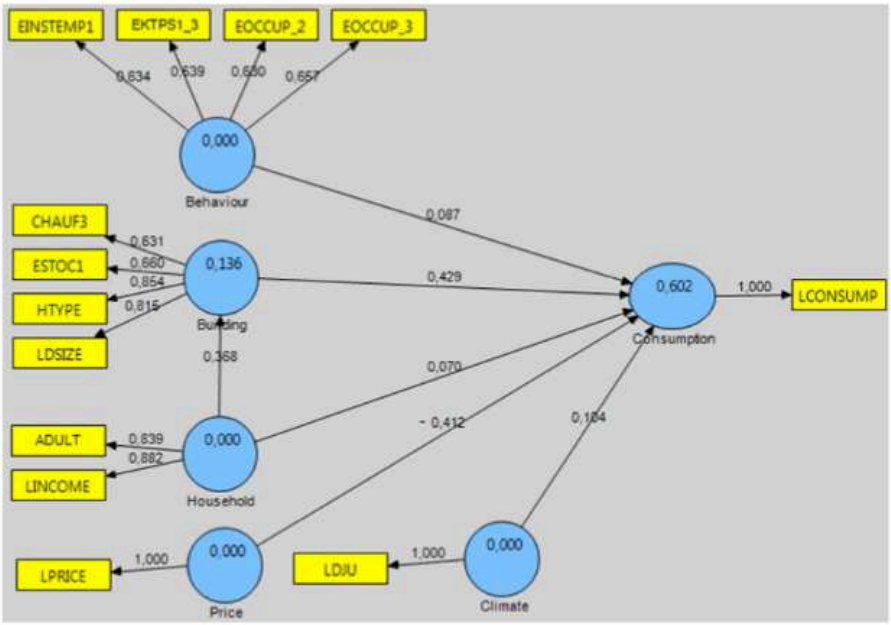


Figure 3: Schéma de causalité liant les variables décrivant le logement (Building), le ménage (Household), le prix de l'énergie (Price), le climat (Climate), les comportements domestiques (Behaviour), et le logarithme de la consommation d'énergie finale (Consumption). La figure est extraite de l'article de (Belaïd, 2017).

L'analyse conditionnelle de la demande

Le modèle d'analyse conditionnelle de la demande (ACD) développé par (Parti et Parti, 1980) est une autre approche de modélisation linéaire dont l'objet est d'étudier une éventuelle différenciation des usages des appareils domestiques. Pour cela, ce modèle de régression introduit un terme d'interaction (voir l'équation ci-dessous) à l'aide de variables d'interférence (caractéristiques des ménages, de l'environnement, des logements). En reprenant les résultats précédents, la CDA s'écrit sous la forme suivante :

$$log\ (CED) = Y = \sum_{i=1}^{N}\sum_{j=1}^{M} b_{ij}(V_j A_i) + \ \varepsilon$$

Où N est le nombre d'appareils considérés, M est le nombre de variables dont nous étudions l'interaction, A_i est une variable indicatrice qui vaut 1 si le ménage possède l'appareil i et 0 sinon. V_j est la variable d'interaction j et b_{ij} est le terme d'interaction. Dans cette approche les données de comportements (équipement, usage) sont particulièrement sensibles. Cette méthode a été utilisée pour modéliser la consommation d'énergie à plusieurs échelles temporelles, de l'année (Papineau et *al.*, 2021 ; Geneviève, 2004) à la journée (Aigner, Sorooshian et Kerwin, 1984). Elle a également été utilisée pour modéliser la consommation du chauffage, du refroidissement (Papineau et *al.*, 2021) ou de l'électricité spécifique (Matsumoto, 2016).

Les deux avantages de cette approche sont (1) la mise en évidence des processus de consommation permise par l'inventaire des appareils et donc des mécanismes par lesquels l'énergie est consommée, et

(2) la différenciation des usages et des consommations associées à des éléments exogènes. Ce peut être des variables caractérisant le ménage, le logement, le quartier. Le modèle de Aigner et *al.* (1984) quantifie ainsi l'interaction entre : les variables de température locale moyenne et mensuelle et celles indiquant la présence d'un système d'air conditionné dans le logement ; le nombre de pièces et la présence d'un système d'air conditionné, et d'un ballon d'eau chaude. Le modèle de Papineau et *al.* (2021) permet d'étudier l'interaction entre le climat local et les variables de date de construction de logement, d'équipement de pompe à chaleur, de rénovations. Le modèle de Matsumoto (2016) est particulier car il étudie les effets différenciés des usages de 12 appareils domestiques selon les caractéristiques socio-économiques des ménages dans le cas du Japon. Les résultats suggèrent que la composition du ménage et le revenu sont des déterminants importants de l'usage des équipements.

Les modèles de régression non-linéaire

Les modèles précédents partagent logiquement une hypothèse de linéarité des effets des variables. Pour autant, cette question est discutée à travers la restriction des sous-ensembles homogènes (de logements, de ménages, de modes de vies) ou par l'introduction de termes d'interaction (produits entre variables). Toutefois, la littérature sur les modèles issus des sciences physiques qui introduisent des paramètres techniques des bâtiments (voir page 36) illustrent la forte non-linéarité de la CED.

Les modèles d'apprentissage non linéaires modélisant la CED annuelle utilisent eux aussi des données issues d'enquêtes, mais peuvent également s'appuyer sur des données issues de capteurs sur des appareils domestiques ou des données historiques de consommation (par exemple issues de compteurs intelligents mesurant la consommation électrique, la consommation électrique horaire agrégée, les usages journaliers ou horaires des équipements domestiques tirés de données déclarées ou de capteurs). Ces modèles font l'objet de nombreux de travaux ayant pour objectif des prédictions de consommations à un horizon temporel court (minute, heure, voire journée) (Ahmad et *al.* 2014). Ils servent de pilotage de systèmes domotiques, à la détection d'anomalies dans la consommation, ou à la planification et au contrôle de réseaux d'énergie (Runge et Zmeureanu, 2019 ; Raza et Khosravi, 2015). Il est intéressant de noter que parmi l'ensemble des travaux ayant modélisé annuellement la consommation énergétique des bâtiments, les travaux ciblant les bâtiments résidentiels sont peu nombreux (Amasyali et El-Gohary, 2018). Nous pouvons présenter cinq différentes approches (ou modèles) de régression non-linéaire :

- Approche par les réseaux de neurones : les réseaux de neurones sont très utilisés dans le champ de la modélisation car ils fournissent généralement une performance d'estimation précise. (Olofsson, Andersson et Östin, 1998) ont combiné un réseau de neurones et une description matérielle des logements pour prédire la consommation d'énergie de chauffage de 10 ménages suédois. Pour estimer la consommation d'énergie électrique de ménages pendant une semaine, sur deux périodes de l'année (été/hiver), et à partir de données d'enquête, (Tso et Yau, 2007) ont comparé la performance de trois modèles. Un premier modèle de régression, un deuxième

basé sur des réseaux de neurones et enfin un dernier basé sur un arbre décisionnel. Dans chacun des modèles, ils analysent d'un côté le poids pris par les variables et, de l'autre, les performances des estimateurs. Les résultats des calculs des trois algorithmes aboutissent à des performances similaires en termes d'estimation. En ce qui concerne les variables sélectionnées, elles sont toutefois différentes selon la période considérée : la possession d'un climatiseur étant par exemple significative pour le modèle de l'été et non sur la période hivernale.

- Approche de régression sur des classes : comme la littérature sur la régression l'a montré, il existe une sensibilité des résultats à l'espace considéré. Pour l'atténuer, il peut être intéressant de segmenter le corpus de données en amont (par exemple en différenciant les maisons individuelles et les appartements (MTES 2017). Une autre méthode peut être de réaliser une classification (non-supervisée) et une régression simultanément en subordonnant le critère de classification à un critère de précision du modèle. Le modèle MARS (en anglais pour « Multivariate adaptive regression splines » (Friedman, 1991) permet d'effectuer ce type d'opération. (Sekhar Roy, Roy, et Balas 2018) ont construit un modèle MARS associant plusieurs sous-modèles de régression multilinéaire pour prédire la consommation de chauffage et de climatisation à partir de paramètres physiques du logement.
- Approche par arbres décisionnels : les modèles par arbres décisionnels sont une famille de modèles mathématiques utilisés pour des tâches de classification et de régression et qui partagent le fait d'obtenir le résultat) la suite d'une série de tests sur les variables (ce qui peut être représenté graphiquement avec un arbre représentant les séquences de décision). Ces modèles ont fréquemment été mobilisés car ils permettent une bonne explicabilité des résultats et sont (en général) simples à calculer.
- Approche par Machines à vecteur de support : les machines à vecteur de support (SVM en anglais pour Support Vector Machine) sont des modèles très utilisés en apprentissage statistique pour des problèmes de classification (resp. un modèle qui doit prédire le niveau d'une variable qualitative) ou de régression (resp. variable quantitative). L'hypothèse de ce modèle est que la projection des données dans un espace de très grande dimension permettra de construire plus facilement une séparation linéaire entre les données. (Chou et Bui, 2014) ont comparé des modèles ANN, SVM, d'arbres décisionnels et un modèle associant un réseau de neurones avec un modèle SVM pour prédire la consommation annuelle simulée de chauffage et de refroidissement de près de 800 logements. Tous les modèles testés offraient une performance intéressante avec plus de 98% de la variance expliquée par 8 variables décrivant des paramètres physiques du logement. De leur côté, (Olu-Ajayi et *al.*, 2022) ont comparé un réseau de neurones artificiels (ANN), un modèle Support Vector Machine (SVM) et un arbre décisionnel (DT). En cherchant à prédire la consommation annuelle par unité de surface de 300 ménages anglais à partir de données physiques des logements et de données météorologiques. Les auteurs montrent que les modèles ANN et SVM ont des performances similaires (variance expliquée

aux alentours de 60%) tandis que le modèle DT est significativement moins performant (R^2 autour de 27%). D'autres travaux réalisé à partir de données réelles chinoises suggèrent que les performances des SVM fournissent de bonnes performances d'estimation relativement aux approches neuronales à partir des description physiques de bâtiments (Qiong Li, Peng Ren, et Qinglin Meng, 2010).

- Modèles d'ensemble. Les modèles dits « d'ensemble » pondèrent les résultats de plusieurs sous-modèles d'estimation, souvent assez simples, qui diffèrent en nature (des algorithmes différents), ou par les données mobilisées (les modèles peuvent être entraînés sur des sous-ensemble de la base de données ou avec un nombre limité de variables).

On observe dans cette rapide revue que les travaux en sciences de la donnée ont exploré l'utilisation de modèles complexes pour estimer la consommation d'énergie domestique. Les résultats montrent que les modèles non linéaires améliorent la performance d'estimation par rapport aux approches linéaires. Par ailleurs, les variables sélectionnées pour prédire la consommation d'énergie sont sensiblement identiques à celles identifiées auparavant dans la littérature. Pour comprendre l'origine des consommations, il semblerait donc qu'il faille chercher à étudier non pas l'effet individuel de chacune des variables, mais celui de leurs interactions à travers l'étude de « classes » (ou *clusters* en anglais).

Les modèles de partitionnement

La construction de groupes est une pratique de modélisation statistique qui consiste à rassembler (on parle également de *clustering*) des lignes d'une base de données selon un critère de similarité. Ce partitionnement peut renvoyer à la recherche de classes *latentes* dans une population, mais il peut aussi servir à échantillonner une base de données afin de faciliter l'analyse ou de permettre une meilleure régression sur chacun des sous-ensembles de données. La littérature montre que ces classifications portent sur des variables diverses (comportements, caractéristiques socio-économiques du ménage, caractéristiques du logement, de l'environnement, données historiques de consommations agrégées ou désagrégées). Aussi, les modèles diffèrent dans la structure de l'algorithme qui peut se composer d'un premier modèle d'extraction de caractéristique (*feature extraction*) et d'un modèle de regroupement. Nous décrivons quelques exemples tirés de la littérature.

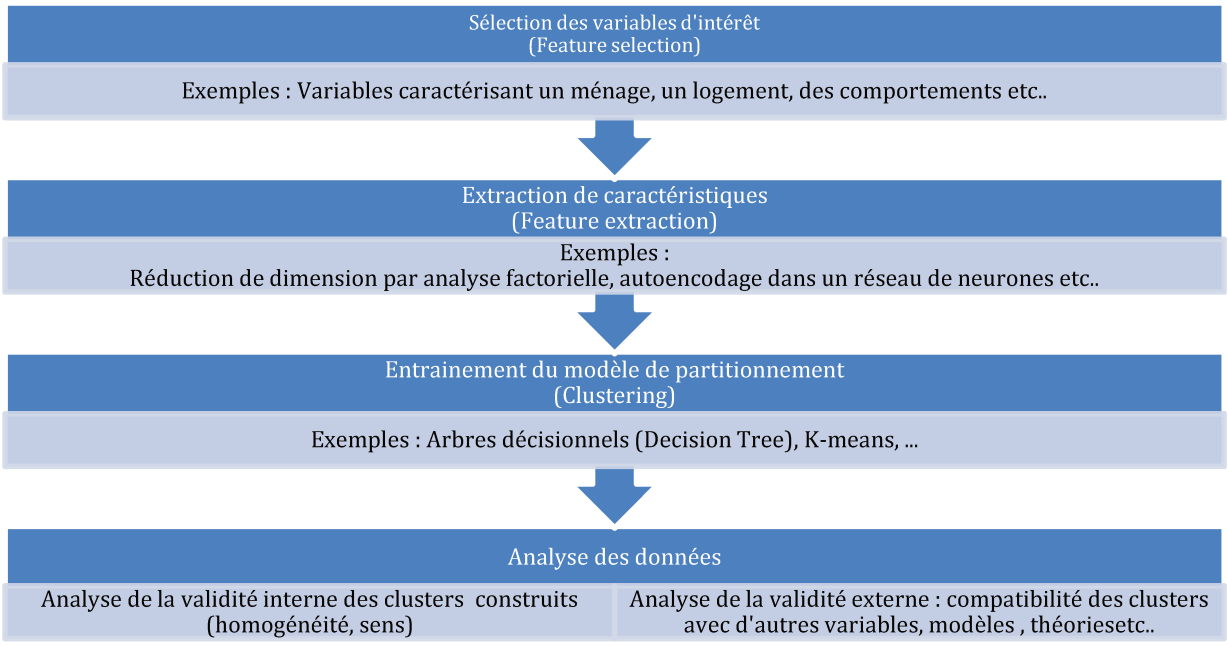


Figure 4 : Méthodologie générale pour le partitionnement de données. Source : Auteur.

On propose ici de reprendre quelques travaux de la littérature ayant réalisé des classifications au sein d'une démarche de modélisation de la CED.

La classification de données de données de consommation d'énergie

Une première approche pour la construction de classe consiste à différencier les consommations d'énergie et étudier les profils de consommateurs associés. Ce type de travail est assez fréquent dans la littérature. Les travaux diffèrent dans le type de données d'énergie (qui peuvent être issues de capteurs ou d'enquêtes), l'échelle spatiale considérée et la stratégie de modélisation. (McLoughlin, Duffy, et Conlon 2015) ont par exemple comparé 3 algorithmes de classification (K-means, K-medoids et un réseau de Kohonen) sur des données de consommations issues de compteurs intelligents pour extraire des profils de consommation et étudier les ménages associés.

La classification de données de comportements domestiques

Les comportements domestiques sont des déterminants majeurs de la CED. Plusieurs travaux se sont intéressés à classifier des comportements énergétiques domestiques, le plus souvent pour construire des archétypes de mode de vie résidentiels. A notre connaissance la première classification de variables de comportements domestiques est celle de Van Raaij en 1983 (van Raaij et Verhallen, 1983) qui a construit 5 types de comportements à partir des pratiques déclarées de chauffage et de ventilation de 145 ménages hollandais. Il a opéré une Analyse par Correspondance Multiple (ACM) sur 17 variables qualitatives, ce qui a permis de construire 6 axes principaux. A partir des scores induits pour chacun des ménages, il opère une dichotomisation (score haut/score bas). Il construit ainsi manuellement 5 types de comportement et mesure la performance d'un modèle logistique de leur prédiction à partir des caractéristiques socio-économiques des ménages et d'attitudes déclarées. Il obtient une performance

moyenne de 64% de prédiction correcte. Cette procédure a été appliquée à plusieurs reprises sur d'autres pays (Ben et Steemers, 2018 ; Ortiz et Bluyssen, 2019), avec d'autres variables de comportement, et sur des ensembles de ménages plus ou moins restreints allant de quelques dizaines à un peu plus d'un millier de ménages (Guerra Santin, 2011 ; Sütterlin, Brunner, et Siegrist, 2011). Les typologies distinguent généralement deux types extrêmes : l'un regroupant des comportements très énergivores et l'autre des comportements peu énergivores voire privatifs.

La classification de données de logements

La classification des logements pour expliquer et prédire la CED est une pratique extrêmement répandue dans la littérature. Elle fournit plusieurs avantages : sur le plan prédictif, la séparation des logements permet de différencier des logements ayant des volumes et des performances thermiques très différentes et garantit ainsi une bonne séparabilité des données de consommation. Ensuite, ces classifications reposent sur des données qui sont aujourd'hui très disponibles, rendant la pratique plus aisée et la méthodologie transférable d'une région à une autre. Parmi les nombreux travaux dans la littérature, nous retiendrons les travaux de référence du projet « Typology Approach for Building Stock Energy Assessment » - TABULA (Ballarini, Corgnati, et Corrado, 2014 ; Loga, Stein, et Diefenbach, 2016) qui ont construit des typologies de bâtiments pour 13 pays européens. Pour la France, 40 types de bâtiments ont été proposés. En termes de méthode, ce projet a utilisé une séparation des logements selon l'âge et le type de logement.

La classification de données d'activités

D'autres travaux se sont intéressés à la classification de données d'activités à l'aide des enquêtes emploi du temps. Celui de (Diao et *al.*, 2017) est un bon exemple des recherches ayant porté sur la mobilisation des comportements pour modéliser la consommation d'énergie. Les auteurs indiquent que les caractéristiques socio-démographiques sont des variables signifiantes de comportements, il est souhaitable d'avoir accès à la logique globale des activités domestiques pour obtenir un véritable modèle explicatif. En utilisant une méthode de partitionnement par K-Modes sur les données d'enquêtes américaines, ils identifient 10 classes de comportements qu'ils utilisent ensuite au sein d'un modèle de simulation *bottom-up* de CED. D'autres travaux discutent de la méthode et des critères de partitionnement. Les travaux les plus avancés étudient le lien entre ces classes de comportement et les caractéristiques du ménage et du logement (Liu, Hu, et Yan 2023).

Les méthodes mixtes

Plusieurs recherches associent les différentes catégories de variables pour construire des classes permettant d'expliquer la CED. Par exemple, (Hache, Leboullenger, et Mignon 2017) ont effectué une étude désagrégée de la CED par classification hiérarchiques des ménages et de leurs factures d'électricité. Dans cette approche originale, les auteurs utilisent une méthode de partitionnement par arbre décisionnel pour construire 38 groupes à partir de l'Enquête Nationale Logement (Figure 5). Un point particulièrement intéressant est que les auteurs ont identifié des variables explicatives différentes selon le type de logement et le type d'énergie. Si le revenu est une variable discriminante parmi tous les groupes de logements, la densité urbaine est uniquement déterminante parmi les appartements chauffés au gaz. Parmi l'ensemble des travaux portant sur la classification des CED, cette étude est originale dans la mesure où elle construit davantage de groupes en différenciant la classification des ménages selon leur type de logement et le mode d'énergie.

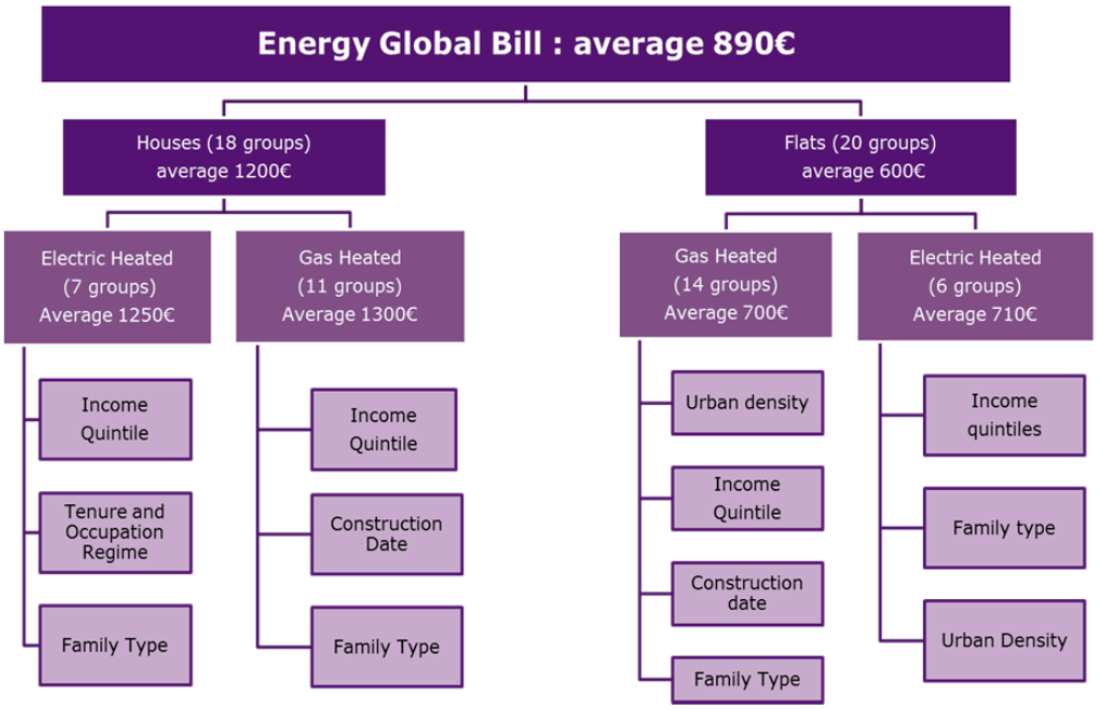


Figure 5 : Typologie des ménages construite par (Hache, 2017).

Dans la littérature, l'article de (Bogin, Kissinger, et Erell 2021) est intéressant car il compare la performance prédictive de deux modèles de classification qui se différencient par les données utilisées. Le premier modèle est basé sur des données d'usage et le second utilise les caractéristiques des ménages et des logements pour construire un partitionnement permettant d'expliquer le niveau de consommation agrégé d'un logement. Les auteurs montrent que la première stratégie est plus efficiente mais soulignent la forte influence de la sélection des données sur les résultats et en soulignant une forte corrélation entre les deux types de variables ils invitent les modélisateurs à considérer un choix de variables adaptées à l'échelle spatiale de l'exercice.

Enfin, il faut signaler les travaux de (Salvó et Piacquadio, 2017). Les auteurs ont construit un modèle de classification de la demande en électricité avec un critère fractal. L'idée qui dirige ce travail est que la demande en énergie est façonnée à différents niveaux par des logiques hétérogènes. L'approche est déjà largement utilisée dans la modélisation spatiale de l'étalement urbain et des inégalités, mais c'est à notre connaissance le seul travail qui cherche à utiliser ce critère géométrique pour construire des classes de consommation d'énergie. En s'intéressant à la distribution spatiale de la demande en électricité, les auteurs différencient deux motifs géographiques de consommations, l'un présent en zone urbain, l'autre en périphérie.

1.2.2 Les modèles « physiques »

Les modélisations dites « physiques » concernent les approches quantitatives qui s'appuient sur des lois physiques afin de modéliser la consommation d'énergie des équipements, des postes de consommation (le chauffage, l'eau chaude sanitaire, l'éclairage, l'électricité pour les usages spécifiques, etc.) ou la consommation totale. Ces modèles diffèrent fondamentalement dans la discrétisation de l'espace du logement, du temps et dans les hypothèses et les choix méthodologiques. Un trait commun à ces modèles est qu'ils reposent sur un bilan énergétique : La formulation d'hypothèses sur le confort souhaité, les paramètres climatiques, le nombre d'équipements et leur intensité d'usage sont mis en balance avec la consommation énergétique des appareils.

On donne deux exemples de bilans énergétiques ci-dessous.

Pour le calcul du besoin énergétique pour l'éclairage. On considère qu'un logement est équipé de 10 lampes, dont la moitié a une puissance nominale de P_1 =10W et l'autre moitié de P_2 =100W. Ces lampes sont utilisées chacune en moyenne 3h par jour. La consommation électrique associée à leur usage journalier est de :

$$\textit{Energie électrique E consommée par les lampes} = \textit{Energie lumineuse et thermique diffusée par les lampes}$$

$$E_{lampes} = (n_1 P_1 + n_2 P_2).\Delta t$$

$$E_{lampes} = (5 * 10 + 5 * 100) * 3$$

$$E_{lampes} = 1{,}7\ kWh$$

Pour le calcul du besoin de chauffage. On considère un logement comportant une seule pièce et où toute la surface (murs, plafond, sol) est exposée à une température extérieure constante de 2°C. Pour maintenir cette pièce à 19°C pendant ΔT =24h, avec une déperdition thermique moyenne de chaleur

de U=1W/m².K (Watts par mètre carré et par degré Kelvin)[16] sur une surface totale extérieure de (par exemple) 65 m² avec une température extérieure moyenne de 2°C il faut une énergie E qui se calcule de la manière suivante :

$$Energie\ E\ fournie\ par\ le\ système\ de\ chauffage\ pendant\ le\ temps\ \Delta t =$$
$$Energie\ thermique\ perdue\ par\ transmission\ à\ l'extérieur\ pendant\ le\ temps\ \Delta t$$

$$E = U.S.\Delta T.\Delta t$$

$$E = 1 * 65 * (19 - 2) * 24 = 26\ kWh$$

Cet encadré est par ailleurs l'occasion de rappeler que si nous parlons de « consommation » d'énergie il s'agit d'un abus de langage. L'énergie utilisée dans les logements n'est en effet que transformée d'une forme en une autre. Dans une lampe, l'énergie électrique est convertie en lumière et en chaleur.

La modélisation de la consommation d'énergie pour le chauffage : la discrétisation du temps et de l'espace dans le logement pour calculer le bilan énergétique

Du fait de son importance prépondérante dans le bilan énergétique du secteur résidentiel, l'estimation du besoin de chauffage des logements a généré un grand intérêt et plusieurs stratégies de modélisation existent. Le tableau 3 propose une typologie des modèles « physiques » selon les différentes discrétisations spatiales et temporelles.

[16] A titre indicatif, un mur de 14cm de béton aura un coefficient de déperdition de près de 4W/m².K , et le même mur complété avec 10cm de laine de verre aura un coefficient d'environ 0.29 W/m².K.

Tableau 3 : Proposition de typologie des stratégies de modélisation "physiques" de la CED selon la discrétisation spatiale et temporelle. Des exemples de modèles sont donnés et décrits dans les paragraphes ci-dessous.

Modèle	Discrétisation spatiale	Discrétisation temporelle	Hypothèses et choix méthodologiques
Modèle par bilans énergétiques moyens Exemple : modèle « 3CL » (Journal Officiel 2021)	Logement	Mois	Paramètres thermiques du logements estimés à partir des caractéristiques observées du bâtiment Bilans énergétiques effectué par mois Usage standardisé du logement selon un scénario de consommation conventionnel (température de chauffe constante de 19°C, occupation à heure fixe).
Modèle dit « RC équivalent » Exemple : modèle « Th-C-Ex » (Journal Officiel 2008)	Logement	Heure	Modèle thermique incluant l'inertie du bâtiment. Les paramètres thermiques du modèle peuvent être construits physiquement ou par apprentissage statistique. Des scénarios de chauffage, d'apport solaire ou d'apports internes peuvent être implémentés.
Modèle multizone Exemple : TRNSYS, EnergyPlus, Comfie-Pléiades (MINES ParisTech 2014),	Pièce du logement	Heure	Maillage du logement en sous-espace thermiques (pièces). Des bilans énergétiques entre les différents espaces sont réalisés à chaque pas de temps de calcul. Le modèle peut être articulé avec des modules modélisant les flux d'air, l'hygrométrie etc.
Modélisation par éléments finis	Système, mur, matériau	Heure	La géométrie du bâtiment est maillée finement. Les paramètres physiques du bâti (température, humidité, vitesse de l'air) sont estimés localement à chaque pas de temps.

Dans cette approche, il s'agit dans un premier temps de réaliser un inventaire exhaustif de paramètres techniques du bâti et de l'environnement liés à l'isolation des murs, la qualité des menuiseries, le climat local. Il est ensuite possible de déduire des paramètres thermiques agrégés qui permettent de déduire une consommation énergétique à l'aide de lois empiriques sur des périodes temporelles larges (mois, année). A titre d'exemple, la méthode dite « 3CL » (pour Calcul des Consommations Conventionnelles) utilisée pour l'élaboration du Diagnostic de Performance Energétique (DPE) mobilise une approche de ce type (Journal Officiel 2021).

Dans la méthode 3CL, on définit alors pour un chauffage constant à 19°C, le besoin de chauffage BV_j au mois j :

$$BV_j = GV.(1 - F_j)$$

Où GV est un coefficient qui modélise les déperditions de l'enveloppe (il est estimé à partir des caractéristiques de l'enveloppe du bâtiment : matériaux, épaisseurs, surfaces) et F_j est la fraction des besoins de chauffage couverts par les apports solaires sur le mois j.

La consommation de chauffage Bch_j est alors estimée chaque mois comme étant :

$$Bch_j = \frac{BV_j.DH_j}{1000} - \frac{Q_{pertes,j}}{1000}$$

Où le coefficient DH_j est le nombre de « degré heure » de chauffage sur le mois j, en considérant un niveau de chauffage à 19°C. Cette valeur est calculée en faisant le produit entre le nombre d'heures de chauffage et l'écart moyen de température entre celle de consigne de 19°C et celle de l'air extérieur (qui dépend de la zone climatique où se situe le logement). Le coefficient $Q_{pertes,j}$ quantifie quant à lui l'ensemble des pertes de chaleur des systèmes de production, de stockage et de distribution d'eau chaude sanitaire à l'intérieur du logement. Le besoin total de chauffage est calculé en calculant la somme de chaque mois. L'intérêt de cette approche réside dans sa capacité à associer une finesse dans la description des caractéristiques du bâti, tout en fournissant une méthode de calcul explicite à travers un système d'équations. En revanche, la description des usages y est faite de manière standardisée : la présence des occupants est modélisée à travers un « scénario conventionnel d'occupation hebdomadaire des logements » avec les hypothèses suivantes :

- Occupation du logement de 0h à 9h et de 17h à 24h avec une période de sommeil allant de 0h à 6h et de 22h à 24h les lundi, mardi, jeudi et vendredi.
- De 0h à 9h et de 13h à 24h avec une période de sommeil allant de 0h à 6h et de 22h à 24h le mercredi.
- De 0h à 24h les samedi et dimanche avec une période de sommeil allant de 0h à 6h et de 22h à 24h.

Ces hypothèses très restrictives, permettent une comparaison « à usage constant » des enveloppes thermiques mais ne rendent pas compte des dynamiques comportementales qui génèrent les consommations. En conséquence, si cette méthode d'estimation peut servir à calculer les effets d'une rénovation sur la consommation d'énergie avec un usage conventionnel, elle ne permet pas d'estimer les consommations réelles qui elles nécessitent la traduction dans le modèle des dynamiques temporelles et spatiales des usages domestiques.

Modélisation avec une paramétrisation R-C

La méthode par modèle « RC équivalent » est une méthode très développée dans la littérature (Swan et Ugursal, 2009) et apporte des éléments permettant d'intégrer ces dynamiques spatiales et temporelles. Le modèle permet de calculer des consommations annuelles de chauffage selon des scénarios d'usages variés, aider au choix de solutions techniques dans le cadre d'une rénovation et pour le design de solutions d'automatisation et de contrôle domotique (X. Li et Wen, 2014). Un certain nombre d'hypothèses sont posées : la température de l'air intérieur est supposée uniforme dans chacune des zones thermiques du modèle (représentant chacune une pièce ou un espace donné) et l'influence de la pression et de l'hygrométrie (le taux d'humidité de l'air) sur les transferts thermiques est négligée. On peut ainsi modéliser le comportement thermique d'un logement à l'aide d'un schéma qui s'appuie sur une analogie avec l'électricité (voir l'exemple donné à la Figure 6). Sur le schéma, les paramètres du modèle (les résistances R1, R2, R3, et la capacité Ce) sont déduits par identification du modèle avec des données expérimentales ou calculées empiriquement (Y. Li et *al.*, 2021). Le logement peut être soit modélisé comme un seul espace thermique (comme c'est le cas sur cet exemple) ou comme un ensemble d'espaces thermiques connectés (Aoun et *al.* 2019) : dans ce dernier cas, le schéma est complexifié par l'ajout de paramètres R et C. Cette modélisation permet de calculer la consommation de chauffage en introduisant des *scénarii* d'usage. Par exemple avec des hypothèses sur l'évolution temporelle de la température intérieure T_i, de la température extérieure T_o, et des apports solaires $F_s \Phi_s$ il est possible de calculer la puissance thermique Q_h nécessaire à chaque instant pour assurer le confort thermique désiré. Il est intéressant de noter que, dans ce modèle, l'inertie du bâtiment est quantifiée à travers le paramètre C_e. L'inertie thermique fait référence à la capacité plus ou moins importante d'un bâtiment à stocker la chaleur puis à la restituer de manière diffuse. Cette capacité permet de lisser les apports de chaleur et de diminuer légèrement les besoins de chauffage (Verbeke et Audenaert 2018).

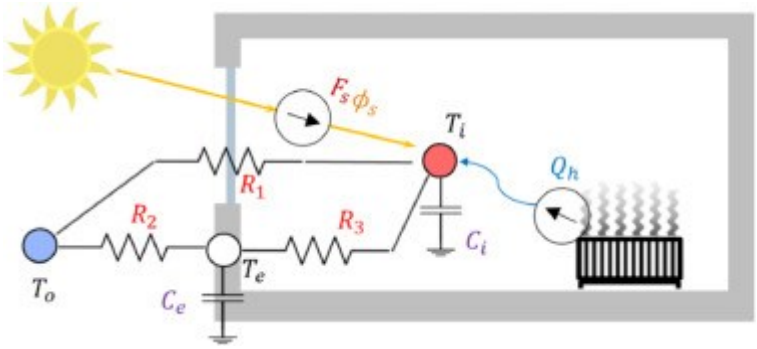


Figure 6 : Représentation d'une pièce chauffée, et du modèle « RC équivalent ». Ti désigne la température moyenne intérieure ; Te la température moyenne de l'enveloppe ; To la température moyenne de l'air extérieur ; Qh le flux de chaleur du chauffage et FsΦs la fraction flux solaire pénétrant à l'intérieur de la pièce. Les résistances R1, R2, et R3 représentent la résistance thermique respective des surfaces vitrées, et la résistance thermique associée au phénomène de convection à l'intérieur et à l'extérieur. La capacité Ce permet de modéliser l'inertie thermique des parois. Le schéma est tiré de la publication de (Wang et Chen 2019).

La littérature montre que ce type de modèle offre de bonnes performances prédictives. Il est utilisé pour le design de systèmes de contrôle et d'optimisation des systèmes de chauffage (Aoun et *al.*, 2019), la modélisation des consommations énergétiques de quartiers, le soutien de programmes de rénovation associés et la simulation de l'intégration des bâtiments dans des réseaux énergétiques tels que des Smart Grids et des réseaux de chaleurs urbains (Reynders, Nuytten,et Saelens 2013).

(Y. Li et *al.*, 2021) soulignent cependant que ces modèles souffrent toutefois de plusieurs défauts : les paramètres R, C sont mal définis et peuvent être soit identifiés à partir des caractéristiques du logement, soit par identification du modèle sur des données collectées de température et de chauffage. Les fondements théoriques sont alors incertains et la complexité de modèles comportant un grand nombre de paramètres R et C peut compliquer l'interprétation.

Modèle thermique multizone

Pour ces raisons, les modèles thermiques qui sont mobilisés dans les logiciels de simulation thermique font appel à des bilans thermiques multizones avec une résolution temporelle de l'heure. Dans cette famille de modèles « physiques », le bâtiment est décomposé en mailles et chacune est associée généralement à une pièce du logement où la température est supposée uniforme à chaque pas de temps de calcul. L'évolution des températures est décrite à l'aide de paramètres de transfert thermique (capacités thermiques de la maille, transfert de chaleur aux frontières de chacune des mailles). Un bilan d'énergie est effectué à chaque pas de temps (généralement une heure). Ce type d'approche est très utilisé dans le domaine du bâtiment à travers des logiciels comme TRNSYS ou EnergyPlus. Le moteur de calcul thermique Comfie développé par (Peuportier et Blanc, 1990) a apporté une méthode mathématique permettant de réduire le temps de calcul de l'ordre de l'heure à quelques minutes. Cependant, ces modèles diffèrent dans leur capacité à intégrer des modules complémentaires permettant de simuler la diffusion de l'air, les apports solaires éventuellement masqués par le voisinage, les usages énergétiques et l'occupation du logement etc.

Enfin, en sachant que l'écoulement de l'air et les apports solaires peuvent exercer une influence déterminante, des modèles ont été construits à partir d'une maquette numérique du logement afin de réaliser un calcul par éléments finis[17]. Ce type de calcul permet d'observer des dynamiques locales. Il est utile principalement au design et à la comparaison de solutions techniques. Nous pouvons citer par exemple le travail de (Pisello et *al.*, 2016) où une modélisation par éléments finis sur l'année permet d'étudier l'écoulement de l'air et les changements de température ambiante sous les combles d'une maison située au centre de l'Italie après l'installation d'un « toit frais ». Ce type de modélisation reste toutefois très gourmand en données descriptives (épaisseur des murs, des fenêtres et matériaux, qualité de la réalisation, variables du climat intérieur et extérieur …) et en capacité de calcul et est de ce fait peu utilisé pour le calcul de la CED des bâtiments.

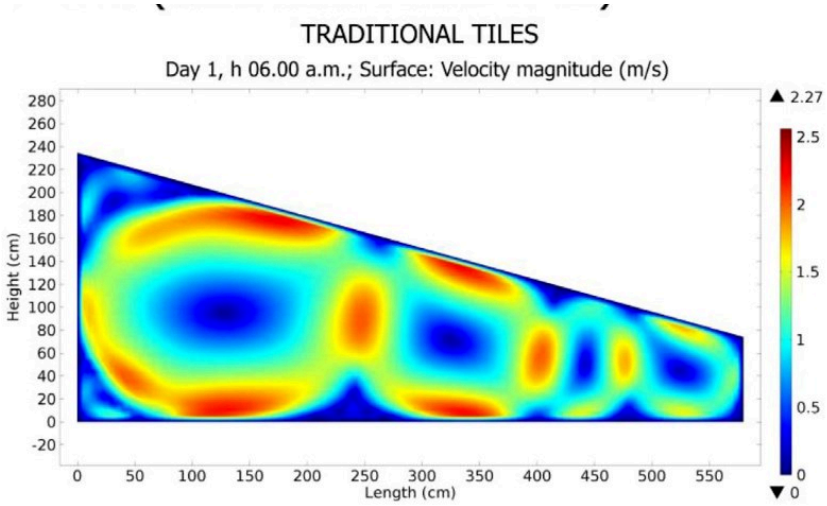


Figure 7: Exemple de résultat de calcul issu d'un modèle de CFD. Tiré de Pisello, 2016.

La modélisation des consommations des autres postes de consommation

Les modèles présentés ci-dessus ont été construits afin de fournir une estimation de la consommation de chauffage. Ils peuvent être articulés à d'autres modèles pour permettre l'estimation des consommations énergétiques des autres postes (éclairage, eau chaude sanitaire, électricité spécifique, refroidissement, ventilation, etc.). Cette association peut être additive (cumul des consommations énergétiques de chaque poste) ou intégrative (les calculs de consommations de chaque poste peuvent être associés dans le modèle). A titre d'exemple de ce dernier type de calcul, nous pouvons citer la prise en compte des pertes de chaleur générées par la production, le stockage et la distribution de l'eau chaude sanitaire (ECS). Ces pertes peuvent tout autant diminuer le besoin de chauffage (elles contribuent à réchauffer les pièces) et augmenter le besoin de refroidissement. Ce type d'interaction entre les modèles

[17] La méthode des éléments finis est un outil mathématique qui permet de résoudre un système d'équations différentielles qui représentent de manière analytique le comportement dynamique de systèmes physiques (ici l'écoulement de l'air dans un logement, la température des murs et de l'air ambiant …). Le milieu matériel est maillé et les équations de transfert de chaleur et de matière sont résolus à l'interface de chacune des cellules, permettant un accès local aux paramètres physiques de température, d'humidité.

est par exemple intégré dans les calculs règlementaires (Journal Officiel 2021) tel que l'impact d'un usage comme l'éclairage du logement sur les pratiques et intensités de chauffage domestique (Richardson, Thomson,et Infield 2008).

Tableau 4 : Exemple de modélisations utilisées pour calculer les consommations énergétiques associées à différents usages. Source : Auteur.

Usage ou poste de consommation *(Modèle)*	Hypothèses de modélisation
Eclairage *(Modèle Th-C-ex)*	**Modèle par équation** Le calcul de la consommation d'éclairage artificiel d'un local est le produit de la puissance d'éclairage artificiel installée (2W/m² dans le modèle Th-C-ex par exemple) par sa durée d'utilisation sur une heure. Les consommations sont sommées sur l'ensembles des zones puis sur l'ensemble de l'année
ECS *(Modèle Th-C-ex)*	**Modèle par équation** La quantité de chaleur nécessaire à la production d'ECS est calculée proportionnellement au volume d'eau chaude, celle-ci étant définie dans les scénarios règlementaires comme proportionnelle à la surface habitée (Journal Officiel 2021).
Eclairage, ECS & Usages d'électricité spécifique *(Modèle de Tanimoto, 2008)*	**Modèle dynamique** Génération d'une séquence d'occupation et d'activités domestiques à partir d'un processus probabiliste. Les paramètres de ce modèle sont calculés par apprentissage statistique sur des données d'enquête « emploi du temps[18] » japonaises. - L'éclairage dans une pièce est considéré allumé que lorsqu'un habitant se trouve dans une pièce mais ne dort pas, et que l'extérieur est suffisamment sombre. - Les appareils électriques ont une consommation moyenne : un réfrigérateur consomme 60W en continu, un micro-onde consomme 200W pendant la durée d'une activité « repas », une télévision consomme 120W durant l'activité « télévision » et 2W le reste du temps etc.

Le tableau ci-dessus (Tableau 4) différencie les travaux qui segmentent la consommation selon le poste, l'usage et l'appareil. Généralement, les auteurs distinguent les modèles qui effectuent des bilans énergétiques à partir de paramètres moyens (résolution d'équation) de ceux qui s'appuient sur la dynamique temporelle des usages très développés depuis le début des années 2000. Le principe des modèles dynamiques consiste à réduire l'information statistique de l'occupation des logements et des activités domestiques des ménages dans une approche probabiliste permettant de générer des séquences d'activités domestiques vraisemblables. Ces modèles ont été appliqués dans différents pays, entre autres: au Japon (Tanimoto, Hagishima, et Sagara 2008), en Grande-Bretagne (Richardson, Thomson,

et Infield, 2008) et McKenna (McKenna et Thomson 2016), en Suède (Widén, Molin, et Ellegård, 2012), aux Etats-Unis (Muratori et al., 2013), en France (Vorger, 2014), et en Belgique (Aerts et al., 2014).

Un facteur différenciant : le choix des hypothèses d'usage

La modélisation des comportements dans les modèles de consommations d'énergie « physiques » peut être réalisée de plusieurs manières. La modélisation des comportements dans les modèles de consommations d'énergie « physiques » peut être produite par différentes approches. Gaetani et *al*. proposent une typologie de ces différentes approches afin de permettre le choix du modèle le plus approprié (Gaetani, Hoes,et Hensen 2016). Ils distinguent dans leur article quatre familles de modélisation des comportements :

- **Les modèles déterministes** consistent en la construction de séquences d'occupation et d'activité dans le logement. Les choix de modélisation (heures d'occupation type d'activité, etc.) reposent généralement sur les recommandations d'instances internationales telles que l'Agence Internationale de l'Energie. Il est par exemple possible de se référer à l'annexe 66 du « Buildings and Community (EBC) Programme » qui fournit des standards pour l'intégration des comportements dans les modèles de simulation du bâtiment (IEA 2017).
- **Les modèles « historiques »**, quant à eux, renvoient à la construction de scénarios (fixes) d'activités et d'occupation reposant sur l'analyse de bases de données. Plus riches que les modèles déterministes et avec une base empirique plus solide, ils souffrent cependant d'une forte dépendance aux données qu'ils utilisent.
- **Les modèles probabilistes** sont une évolution des modèles « historiques » dans le sens où ils permettent, à partir de bases de données, de construire des modèles probabilistes (comme un modèle de chaînes de Markov). Par construction, ces modèles permettent de générer des scénarios vraisemblables avec une résolution fine, les présences et les activités étant déclenchées selon des lois de probabilité.
- **Les modèles multi-agents** sont une dernière famille qui porte sur les décisions individuelles et les interactions entre individus, systèmes et bâtiments. Il s'agit des modèles les plus complexes à développer au sens où ils requièrent, pour les calibrer, un grand nombre de paramètres, d'hypothèses et de données empiriques.

En observant la dynamique des recherches dans ce champ, Gaetani et *al*. observent une massification des approches stochastiques et multi-agents qui permettent d'étudier et d'intégrer la diversité des usages domestiques ainsi que leur caractère « aléatoire ». Les problématiques contemporaines de ce champ relèvent de l'intégration de l'enchevêtrement des usages et de la collaboration des habitants dans les activités.

1.3 Les modèles qualitatifs

La modélisation qualitative de la CED peut apparaître paradoxale puisque dans ce type de recherche, les auteurs soulignent régulièrement la difficulté à estimer et prédire les consommations énergétiques. Les recherches menées dans les disciplines rattachées aux sciences humaines se détachent d'un formalisme mathématique parfois contraignant, et valorisent d'autres outils comme l'enquête de terrain, l'entretien et l'observation. Cette alternative scientifique engage également d'autres paradigmes de recherche non plus seulement positiviste ou post-positiviste mais plutôt compréhensifs, et permet de dégager des résultats très éclairants sur les dynamiques de la CED. Un état de l'art des contributions issues de l'économie, de la psychologie, de l'anthropologie et de la sociologie est proposé. Pour faciliter la lecture, un court encadré présente quatre paradigmes importants qui dirigent de nombreux travaux de recherche.

Paradigme Positiviste

Le paradigme positiviste est une position philosophique développée par Auguste Comte au XIX^{e} siècle qui dans sa posture par rapport à la réalité (ontologie) est d'abord réaliste : seuls sont dignes d'intérêt les objets observables sur lesquels il devient possible d'étudier leurs relations et d'établir des lois, indépendamment du chercheur et de considérations métaphysiques. Sur cette base, le chercheur pourra sélectionner des outils adaptés pour les observer, les comprendre et *in fine* découvrir des lois causales. Dans son rapport à la connaissance produite (épistémologie), le chercheur est objectiviste : il est en capacité de distinguer ce qui relève de sa personnalité, ses valeurs, de la connaissance qu'il produit. En termes de méthode, l'approche positiviste repose également sur la logique disjonctive. Les critères de validité scientifique sont alors la vérifiabilité, la confirmabilité et la réfutabilité. En synthèse, le positivisme amène dans la recherche en sciences sociales la production de lois dans des contextes sociaux contrôlés par le chercheur. Le positivisme, issu d'abord des recherches en sciences naturelles a largement influencé la construction des premières sciences sociales et s'est initialement imposé comme la « seule manière » de produire des connaissances scientifiques sur le réel. Dans ce paradigme, la connaissance scientifique correspond à une architecture cohérente de savoirs vérifiables et de lois de causes à effets. Il est important de rappeler que même si le positivisme ou aujourd'hui le post-positivisme sont très rattachés aux méthodes quantitatives, les recherches qualitatives peuvent tout à fait y être associées.

Paradigme Post-positiviste

La posture philosophique post-positiviste est née dans la 2^{e} moitié du 20^{e} siècle avec l'émergence des critiques du paradigme positiviste. Karl Popper en est une figure emblématique. On citera par exemple les découvertes en physique quantique qui ont chamboulé « le bloc cohérent et stable » des savoirs préexistants, et l'observation de la construction sociale de la science. Le courant post-positiviste

propose un déplacement d'une ontologie réaliste « naïve » vers une ontologie réaliste mais critique. La recherche post-positiviste reconnaît la présence de biais dans la recherche de la connaissance (formulation du problème, choix des variables, de la méthode de traitement des données par exemple) qui ne pourra être qu'imparfaite. Ce courant ne nie pas l'existence des objets étudiés dans la nature, mais relativise la portée des études scientifiques en indiquant qu'un résultat « positif » diminue la probabilité que des études complémentaires puissent les infirmer. En termes de méthodes, le post-positivisme reconnaît une place importante aux méthodes qualitatives et quantitatives.

Paradigme Constructiviste

La posture constructiviste consiste à penser que notre représentation (concepts, modèles, compréhensions) de la réalité (c'est-à-dire l'ensemble des objets sociaux et naturels ; les pommes, les partis politiques, les atomes ...) est le produit de l'esprit humain et non une image fidèle du réel lui-même. Elle s'oppose à une vision réaliste des choses. C'est Kant qui a initié cette école de pensée, qui se résume très bien avec les mots de Gaston Bachelard (1884–1962) qui parle de l'importance de la formulation de la question dans le processus de recherche : « Et, quoi qu'on en dise, dans la vie scientifique, les problèmes ne se posent pas d'eux-mêmes. C'est précisément ce sens du problème qui donne la marque du véritable esprit scientifique. Pour un esprit scientifique, toute connaissance est une réponse à une question. S'il n'y a pas eu de question, il ne peut y avoir de connaissance scientifique. Rien ne va de soi. Rien n'est donné. Tout est construit » (Bachelard, *La Formation de l'esprit scientifique,* 1938). Le paradigme constructivisme s'inscrit dans une posture interprétative où la connaissance est construite dans l'interaction entre le chercheur et le phénomène. La validité des théories et des modèles produits vient de l'adéquation entre le modèle de l'expérience avec l'expérience. Le constructivisme s'est largement développé dans les milieux de recherche en sciences sociales depuis les années 1980.

Paradigme Critique

La posture critique reconnaît elle aussi une vision subjective de la réalité du chercheur et des personnes interrogées mais donne une importance particulière à ces témoignages pour révéler les enjeux de pouvoir, de domination ou d'oppression dans les contextes sociaux étudiés. L'objectif poursuivi dans ce type d'approche est de produire des connaissances qui ne participent pas à la reproduction des inégalités ou de rapports de domination. Elle s'oppose en particulier aux recherches positivistes puisque ces dernières ne permettent notamment pas de rendre compte des réalités vécues par des minorités, et tolère mal la remise en cause des structures (théories, variables, catégorisations) qui sont identifiées comme immuables. La validité de ce type de recherche vient donc du fait de la capacité des résultats à émanciper et donner du pouvoir à la communauté de recherche.

1.3.1 Approches économiques et psychologiques

Dans ce paragraphe nous regroupons un ensemble de contributions scientifiques apportées par la psychologie et l'économie. Si les objets, les méthodes, les théories ne sont pas les mêmes, nous les regroupons ici plutôt par commodité de présentation. Les approches économiques de modélisation explicative de la CED ont déjà été partiellement abordées dans la partie précédente puisqu'un nombre important de travaux en sciences économiques mobilisent la régression multilinéaire ou des systèmes d'équations structurelles pour produire des estimations quantifiées de la CED. Il convient cependant de compléter les contributions de ces disciplines en présentant les hypothèses comportementales des différentes théories issues des sciences économiques et de la psychologie. Cette revue de littérature des théories économiques et psychologiques se nourrit de l'excellente revue de la littérature effectuée par Heydarian (Heydarian et *al.*, 2020).

La rationalité (limitée) des comportements domestiques

La théorie de l'acteur rationnel est un paradigme individualiste issu des sciences économiques. Les comportements observés sont, selon cette théorie, associé à des logiques absolues (température de confort, utilité des investissements dans des appareils ou des rénovations etc.). Les travaux dans le champ économique ont cependant montré que les individus avaient le plus souvent une connaissance imparfaite des choix et des contraintes introduisant des biais cognitifs. Dans ces conditions, les choix des individus sont sous-optimaux et dépendant du niveau d'information, des ressources et de l'environnement dans lequel ils évoluent.

Théorie du comportement planifié

Une autre compréhension des comportements dans le champ économique est celle proposée par Icek Ajzen (Ajzen, 1985). Il postule que les comportements des individus sont planifiés. Pour observer un comportement, il est nécessaire de rassembler trois facteurs : l'attitude (l'individu se déclare favorable vis-à-vis du comportement) ; les normes sociales (le comportement se conforme aux normes sociales dominantes) ; le contrôle perçu (une croyance de l'individu dans sa capacité à obtenir ses besoins). Ce modèle assez général a été très utilisé dans la conception de programmes d'action publiques.

D'autres (nombreuses) théories et modèles psychologiques existent :

- Le modèle d'activation de la norme (« Norm Activation Model » – NAM) proposé par (Schwartz 1970) a été valorisé pour expliquer les comportent pro-sociaux et en particulier l'adoption de pratiques de régulation du chauffage.
- La théorie des besoins, des opportunités et des compétences (« Needs, Opportunities, Abilities » - NOA) formulée par Gatersleben et Vlek (Vlek 1998). Dans cette perspective, on présuppose que l'individu a des besoins. Les opportunités et les compétences des individus sont respectivement des facteurs motivant et limitant : « les besoins associés aux opportunités

constituent la motivation d'achat, tandis que les opportunités associées aux compétences constituent le contrôle comportemental » (traduction par l'auteur).

On notera que les travaux d'économétrie ont largement alimenté la recherche sur ces théories en étudiant et quantifiant l'élasticité de la CED au niveau de l'information (à travers le niveau d'éducation entre autres), ou au prix de l'énergie. Par exemple, (Alberini, Gans,et Velez-Lopez 2011) ont montré que cette élasticité postulée change peu avec le revenu et se situe entre 0,2 et 1,6. (Belaïd 2017) montre à partir de données françaises que le prix de l'énergie est un déterminant significatif de la CED et que son élasticité se situe entre 0,56 to 0,92.

1.3.2 Approches anthropologiques et sociologiques

> *Economic models "usually make strong assumptions about price responses that probably distort the cognitive processes that mediate those responses" (Stern, 1986)*

Des critiques concernant les approches économiques

Des critiques ont été émises sur les modélisations basées sur la rationalité des acteurs. Celles-ci sont liée d'une part à la formulation théorique et d'autre part à la performance empirique. Les articles de (Stern, 1986) et (Hackett et Lutzenhiser 1991) nous paraissent intéressant pour synthétiser les critiques et les améliorations que les modèles sociologiques entendent apporter.

Hackett et Lutzenhiser soulignent que les modélisations économiques donnent une part prépondérante à une variable supposée rationnelle (l'utilité, le revenu, le prix marginal, etc.), sans intégrer le rôle des processus sociaux ou alors de façon très indirecte. Ainsi, les routines sont prises en compte comme des variables de contexte alors qu'elles sont centrales dans l'étude sociologique des comportements et des consommations d'énergie des ménages. De plus, le logement est perçu comme un espace neutre et moyenné, équivalent aux autres logements possédant des caractéristiques physiques semblables.

Un regard particulier sur les travaux d'économétrie est apporté par Alain Desrosières (Desrosières 1995), qui les a comparé aux autres sciences sociales. En observant une divergence sur les objets d'études (les variables pour les premiers, des classes pour les seconds), il montre l'existence d'une divergence d'intérêt. Selon lui, l'économétrie a vocation à décrire un ensemble d'effets en articulant des « variables » dans un modèle de cause à effets reproduisant l'hypothèse d'un processus déterministe. Alain Desrosières note cependant qu'un effort particulier est fourni pour assurer la cohérence interne du modèle (par exemple en testant la significativité des coefficients) plutôt que sur le questionnement du réalisme du modèle statistique construit. En fait, comme le relèvent également Hackett et Lutzenhiser, c'est la question de la place des variables d'intérêt dans les processus qui est soulevée : « le rôle du prix n'est pas clairement établi, (il) peut être un produit du comportement qu'il est censé prédire et sa corrélation avec la consommation peut être fallacieuse » (p. 460, traduction de l'auteur). Ce qui amène ces auteurs à proposer un élargissement du regard du chercheur vers les identités et les rôles sociaux des

individus et des ménages ainsi que vers les pratiques sociales (même si la terminologie et la conceptualisation de celles-ci sont encore peu développées à cette période). Les auteurs prennent l'exemple de la transition de la tarification forfaitaire à la tarification individuelle pour illustrer la compatibilité et l'apport des approches sociologiques : « L'argument est que les coûts agrégés (et peut-être préétablis ou "fixes") sont intrinsèquement moins douloureux que les coûts désagrégés ou variables, même si les quantités totales concernées sont les mêmes. Une version plus sociale de cet argument consisterait à dire qu'il est plus facile pour un individu d'assumer un coût partagé qu'un coût personnel ou privé du même montant » (p. 461).

Ces lectures permettent d'entrevoir que l'intégration des coûts et des informations, des attitudes et des comportements dans un schéma explicatif nécessite de les réintégrer dans des espaces sociaux. Les travaux développés en anthropologie et en sociologie participent alors à partir des cas d'études à identifier de nouvelles variables, de nouveaux concepts et de nouveaux schémas explicatifs permettant de comprendre la CED[19].

Les prémisses de l'étude des comportements domestiques : habiter dans un « monde moderne »

> *« We cannot explain private energy use without understanding the meaning of homes » (Aune 2007)*

Les travaux de Nicole Haumont sur l'« l'habiter »

Dans les années 1960, de nouveaux travaux ont étudié les effets de l'urbanisation sur les modes de vie, en particulier dans les espaces domestiques. Nicole Haumont a développé une analyse du lien entre le logement et les pratiques domestiques. Elle aborde l'espace domestique au-delà de la satisfaction des besoins « partiels » que sont le repos, l'hygiène, l'éducation des enfants, l'alimentation, le loisir (Haumont 1968). L'étude des pratiques et des discours des habitants lui permet d'identifier une logique à un niveau supérieur, dépassant la logique fonctionnelle. Dans ses travaux la chercheuse montre que l'espace du logement n'est plus neutre, ni homogène, les habitants y attribuant des fonctions (« le coin chaud », « la zone de sieste », « l'endroit pour cuisiner », …).

Ces travaux soulignent le fait que les pratiques habitantes sont liées avec un modèle culturel intégré par le ménage, et un habitat. Marion Ségaud rappelle toutefois que cette compréhension de l'habiter ne permet pas une description exhaustive des manières d'habiter.

> *« L'habiter est un fait anthropologique, c'est-à-dire qu'il concerne toute l'espèce humaine. Il s'exprime à travers les activités pratiques dans les objets meubles et immeubles ; il se saisit par l'observation et le langage (la parole de l'habitant). [...] Si l'habiter est un phénomène général, il y a autant de manières d'habiter que d'individus. Dans nos sociétés c'est la conjonction entre un lieu et un individu singulier qui fonde l'habiter. » (Segaud, 2009)*

[19] Cette partie de l'état de l'art est largement inspiré des travaux de Hélène Subrémon et Marguerite Bonnin qui ont réalisé un exercice de synthèse complet, pédagogique de cette famille des travaux : cette note tient lieu de remerciement et de reconnaissance de leur travail.

La définition difficile du « confort »

Le confort est un mot très présent dans la littérature sur l'utilisation de l'énergie dans l'espace domestique. Il renvoie cependant à des concepts différents selon les champs scientifiques et les échelles spatiales (du logement au territoire). Pour les ingénieurs, le confort est souvent réduit au confort thermique (Taleghani et *al.*, 2013), dont la caractérisation est discutée en termes d'indicateurs (la température de chauffage, la fréquence d'ouverture des fenêtres). Pour des chercheurs en sociologie comme Monnier (Monnier 1982), le confort est un assemblage de gestes et d'équipements qui participent à manifester l'appartenance d'un ménage à une culture. Il argumente cette thèse en étudiant le lien entre caractéristiques des ménages (revenus, catégorie socio-professionnelle) avec leurs comportements et leurs revenus. Pour d'autres, comme Michel de Certeau (Certeau, Jameson,et Lovitt 1980), le confort doit être compris à l'échelle locale (celle des ménages dans leur logement) et en relation avec leur passé (éducation, expériences résidentielles) et le contexte présent. Il rejoint en ce sens la position de Lutzenhiser pour qui la « normalité » est relative à un contexte social et matériel. Bonnin résume cette définition en parlant du confort comme d'un « projet évolutif », partiellement construit lors de la socialisation primaire des individus (parents, familles, amis), constamment amendé par les contraintes et les projets liés aux dimensions matérielles, sociale des individus et des ménages (Bonnin 2016).

Le concept « d'appropriation »

D'autres recherches se sont intéressées à l'interaction entre les éléments matériels (le logement, les équipements, le décor), les usages et les consommations d'énergie. Dès lors que l'on ne considère plus l'espace matériel du logement comme un simple cadre ou une scène pour la réalisation des comportements domestiques, il devient possible d'envisager une interaction (une modification réciproque signifiée par le terme « d'appropriation ») entre des usages et des représentations d'un côté et des « systèmes domestiques » de l'autre (Bonnin, 2016). L'analyse des comportements domestiques (les pratiques mises en relation avec les dimensions matérielles et cognitives) permet l'identification des logiques habitantes et des modèles culturels (Subrémon, 2009 ; Bonnin, 2016).

Paradoxalement, le processus d'appropriation est encore plus visible lorsqu'il ne peut pas être mis en œuvre. Par exemple, quand un bailleur refuse au locataire des travaux de décoration ; quand une copropriété empêche de modifier des systèmes d'ouverture et de ventilation ; ou, plus généralement, lorsque des pratiques relevant de modèles culturels ne sont pas conformes aux normes dominantes (par exemple, l'usage à titre privatif des espaces collectifs). Il en résulte que les contraintes d'appropriation, notamment dans le domaine énergétique, sont variables selon les statuts d'occupation. Ainsi, dans le parc locatif, et notamment dans le parc social, le chauffage est souvent collectif, géré et entretenu par le bailleur. Les factures ne sont pas indexées sur les consommations individuelles des ménages, ce qui provoque une distanciation et une non-appropriation des températures de chauffe. Par ailleurs les

contraintes économiques des ménages pèsent sur leur capacité à financer cette appropriation : le prix de l'énergie, des éléments de décoration, des équipements de loisirs, d'hygiène (..) sont des facteurs limitants pour les habitants.

Un autre cas de figure de la rétroaction des systèmes matériels sur les ménages peut être illustré lorsque l'apparition d'un nouvel appareil électroménager (le lave-linge, le sèche-linge, le micro-onde, le lave-vaisselle, …) engendre une facilitation et de nouvelles contraintes du travail domestique (essentiellement des femmes) tout en augmentant la représentation dominante des normes de confort (Cowan 1985). Dans un autre registre, tout autant paradoxal, le congélateur a permis l'achat en grandes quantités de nourriture par les couches précaires en leur permettant de réaliser des économies alimentaires, tout en nécessitant un coût d'investissement et une augmentation de la consommation d'électricité (environ 300 kWh/an selon l'ADEME soit près de 60€/an/ménage).

Comportements identitaires et investissement du logement

Nous avons vu que l'appropriation était une notion clé pour comprendre la façon dont un modèle culturel peut être mis en œuvre par les occupants du logement. Des contraintes matérielles, organisationnelles ou économiques pèsent également sur les marges de manœuvre de l'appropriation. Toutefois, ce processus peut prendre plusieurs formes selon les logiques identitaires des habitants (Bonnin, 2016). Ainsi, selon Bonnin, « la perception que les ménages ont de leur propre logement conduit pour eux à la construction d'une image de celui-ci, et, selon le degré d'adéquation de celle-ci à leurs attentes, se produit un processus d'appropriation qui prend différentes formes » (p. 295). Dans sa thèse, elle identifie trois types d'investissement : l'investissement à perte (coûteuse pour le ménage mais signe d'une adéquation entre forme et usage et, en conséquence, d'une image positive du logement) ; l'investissement d'entretien (appropriation moyenne du logement par le ménage, signe d'une instabilité du ménage dans son logement) ; le sous-investissement (appropriation très faible pour le ménage car les efforts d'investissements sont trop coûteux pour faire corresponde la forme du logement à ses besoins et son identité).

Ses résultats soulignent l'importance du lien entre le ménage et son logement dans l'explication des comportements domestiques, perçus comme des processus plus ou moins aboutis d'appropriation. Ainsi, un ménage qui occupe un logement de mauvaise qualité aura un investissement domestique peu élevé (achat d'équipements, décoration, …), dans la mesure où son identité habitante symbolise sa place dans la hiérarchie sociale. En conséquence, il apparaît que les liens entre l'identité habitante et la représentation sociale de la qualité d'un bien matériel, tel que le logement, agit dans la construction des gestes et des consommations d'énergie domestique.

Ces travaux mettent en évidence l'importance, à l'échelle du logement, de la dimension matérielle, des normes sociales, des structures de ménages et des contextes règlementaires pour comprendre les comportements et les consommations énergétiques. Mais ces questions peuvent également être appréhendées à des échelles spatiales plus étendues et dans d'autres cadres théoriques.

Une approche relationnelle : la théorie de l'acteur réseau

La théorie de l'acteur réseau est une perspective sociologique qui met l'accent sur le caractère relationnel entre des humains et des « non humains » qui interagissent au sein d'un réseau. Ce cadre théorique a été initialement développé dans les années 1980 par Michel Callon, Bruno Latour et Madeleine Akrich pour expliciter les dynamiques (relationnelles) opérant dans la création de la connaissance scientifique. (Latour, 2007). Ce cadre théorique a été prolongé pour étudier les comportements et les consommations d'énergie dans l'espace domestique. Il permet de concevoir les systèmes domestiques comme des acteurs d'un réseau véhiculant des scénarios, lesquels sont ensuite traduits par les occupants des logements. Dès lors, il n'existe plus de hiérarchie entre les humains et les éléments non-humains du logement : chacun est considéré comme un facteur d'égale importance dans la compréhension d'un système domestique formé. Il fournit les bases d'une compréhension des mécanismes de contraintes, d'opportunités et de modification des besoins que génèrent les systèmes techniques domestiques. Il est également très utile pour l'étude de l'appropriation de ces systèmes par des habitants. Ce cadre a par exemple été utilisé par Chiu (Chiu et *al.*, 2014) dans leur enquête auprès de dix ménages occupant des immeubles sur lesquels des travaux de rénovation ont été effectués. Les auteurs soulignent le besoin de considérer « l'adaptation interactive » des usagers dans un environnement sociotechnique pour parvenir à une meilleure évaluation des performances d'une rénovation thermique.

Une approche poststructuraliste des comportements : la théorie des pratiques

Un autre champ important des travaux de recherche sur les consommations d'énergie domestiques est celui traitant de la théorie des pratiques. Dubuisson Quellier et Plessz (Dubuisson-Quellieret Plessz 2013) reprenant la définition de Reckwitz (Reckwitz 2002), définissent les pratiques comme des « types de comportement routinisés qui consistent en plusieurs éléments interconnectés entre eux : des formes d'activités corporelles, des formes d'activités mentales, des "choses" et leur usage, des connaissances de base constituées de compréhension, savoir-faire, états émotionnels et motivations » (traduction reprise de l'article de Dubuisson Quellier et Plessz). Mise en perspective avec les travaux de Bourdieu, Giddens et Shove (Shove, Pantzar, et Watson 2012)., cette définition permet, en analogie avec les théories physiques des champs et de la diffusion, d'étudier les dynamiques spatiales et temporelles de gestes routiniers, comme les pratiques associées à l'entretien et l'hygiène dans le logement, à la cuisson, à l'éclairage, au chauffage.

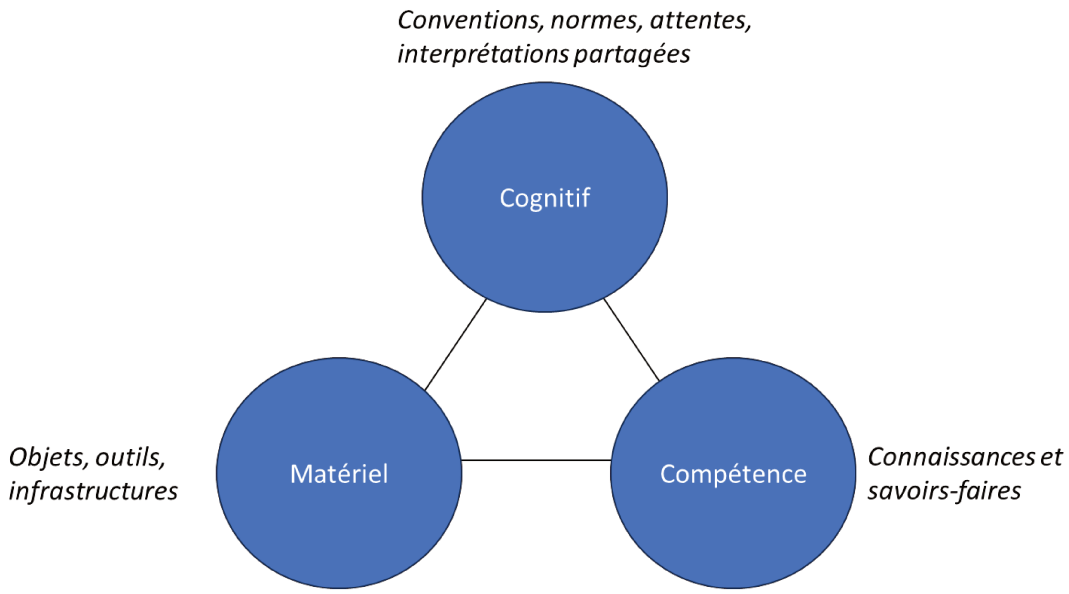


Figure 8 : Les 3 pôles permettant de caractériser une pratique sociale. Adapté de (Shove, Pantzar,et Watson 2012).

En effet, la théorie des champs esquissée par Bourdieu indique que les observations individuelles doivent être perçues et comprises dans un périmètre plus large ainsi que le formule Jon Levi Martin dans son livre sur la théorie des champs (Martin 2003).

> *As Bourdieu says, "To think in terms of field demands a conversion of the whole ordinary vision of the social world which fastens only on visible things [i.e., the individual and the group] (..). In fact, just as the Newtonian theory of gravitation could only be constructed against Cartesian realism which wanted to recognize no mode of action other than collision, direct contact, the notion of field presupposes a break with the realist representation which leads us to reduce the effect of the environment to the effect of direct action as actualized during an interaction." (Martin, 2003)*

Dans cette théorie, les pratiques sont définies comme étant des entités sociales qui « recrutent » des individus et qui sont observables lors de leur « réalisation » dans ces contextes locaux. Cette perspective est d'autant plus intéressante de notre point de vue qu'elle permet d'articuler les dimensions sociales et individuelles des gestes de consommation domestiques : elle consiste, comme le rappellent Dubuisson Quellier et Plessz, en une tentative de théorisation sociologique poststructuraliste (Figure 9), c'est-à-dire qu'elle appréhende les comportements non plus exclusivement sur la base de variables individuelles ou sociales, mais d'une association entre ces deux éléments.

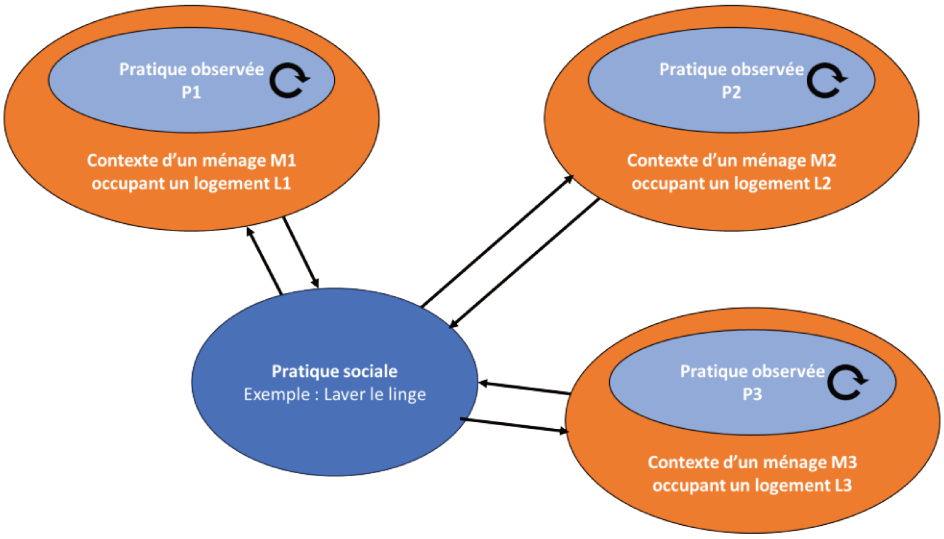


Figure 9 : Schématisation de la dynamique des pratiques sociales. La pratique est définie par l'observation conjointe de comportements, matériels, discours locaux. La pratique s'impose aux ménages mais est réappropriée et redéfinie par eux dans les contextes locaux. Source : Auteur.

Ce cadre théorique se pose en alternative aux approches économiques rationnelles et aux approches psychologiques comportementales. Il a été mobilisé dans de nombreux travaux sur les pratiques d'hygiène, de chauffage, d'usage de l'eau, de mobilité etc. A titre d'exemple, Elizabeth Shove a dirigé de nombreux travaux portant sur les pratiques domestiques, en particulier l'hygiène (Shove 2003).

Les approches culturelles : les travaux sur les modes de vie

Au sein de ces travaux, ceux portant sur les modèles culturels (« Cultural studies ») et les pratiques énergétiques des espaces domestiques abordent les comportements et les consommations comme étant issus d'une interaction entre des caractéristiques d'un environnement social d'une part et matériel d'autre part (caractéristiques physiques et symboliques du ménage et du logement, normes, règles, prix). Un exemple de cadre théorique, inspiré des travaux de Lutzenhiser (Lutzenhiser 1992), est celui du « Energy Culture Framework » qui a été développé dans les années 2010 par Janet Stephenson (Stephenson et *al.*, 2010) comme support théorique pour la réalisation de travaux interdisciplinaires.

Tout en tenant compte de la dimension sociale des comportements, ils centrent leur analyse sur une approche systémique intégrant tout à la fois les comportements domestiques, les matériels, les facteurs externes au logement et les normes. Ce cadre considère trois catégories d'éléments :

- les Normes c'est-à-dire « des croyances partagées sur la manière dont les individus devraient se comporter dans un contexte donné, des attentes et/ou des aspirations concernant ce que nous faisons et ce que nous avons » ;
- les Matériels c'est-à-dire « les technologies et les infrastructures qui influencent l'utilisation de l'énergie, qu'elles soient fonctionnelles ou symboliques » ;

- les Pratiques qui « représentent des "actions habituelles ou coutumières". (…) Elles comprennent à la fois les comportements de routine et les actions peu fréquentes qui sont communes au sein d'un groupe social ».

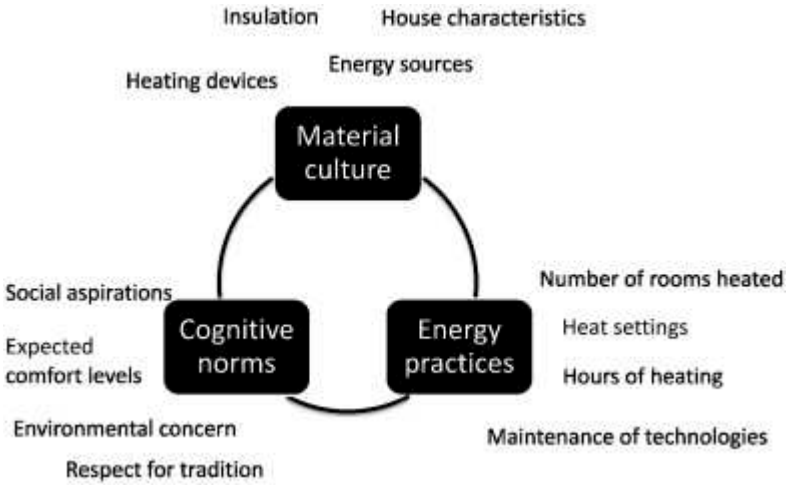


Figure 10 : Le cadre "Energy Culture Framework" est basé sur l'articulation de 3 dimensions (matériel, cognitif, pratiques). Ces 3 dimensions participent à définir une "culture énergétique". Sur l'image la dimension matérielle est définie par le niveau d'isolation, les sources d'énergie etc. Chacune de ces dimensions est soumise à l'influence des facteurs externes (ex : prix des travaux de rénovation, niveau de subvention et d'information etc.). Image reprise de l'article de (Stephenson et al. 2010).

De nombreux exemples de travaux mobilisant une analyse croisée de ces 3 dimensions et faisant référence à une approche « culturelle » peuvent être trouvées dans la littérature.

Willhite (Wilhite et *al.*, 1996) ont, pour leur part, mené une étude comparative des pratiques de chauffage, d'usage d'eau chaude et d'éclairage entre des ménages norvégiens et japonais. En étudiant les liens entre les pratiques quotidiennes et les facteurs culturels et économiques, les auteurs soulignent la surdétermination de la dimension culturelle (bain dans le quotidien au Japon, pratiques de chauffage de la maison en Norvège).

(Rau et *al.*, 2020) ont, quant à eux, suivi 20 ménages lors d'une opération de rénovation et analysé l'évolution des dimensions cognitives, matérielles, comportementales ainsi que l'évolution des consommations de chauffage et d'eau chaude. Ils montrent que le cadre théorique permet d'expliquer les évolutions croisées des paramètres (matériel, cognitif, pratiques, énergie). Par exemple, les ménages ayant manifesté un sentiment accru de satisfaction du confort thermique après rénovation sont aussi les moins aisés dont les consommations en gaz sont les moins élevées avant rénovation. Les travaux de rénovation peuvent alors être compris comme le levier matériel ayant permis la réalisation d'une norme de confort jusqu'alors insatisfaite. Les auteurs montrent cependant que les attentes en termes de température « idéale » sont très variables d'un ménage à l'autre.

Enfin, les travaux de Marguerite Bonnin et d'Hélène Subrémon, déjà évoqués, s'inscrivent dans ce champ de recherche. En dépassant la vision rationnelle pour expliquer l'émergence de comportements

et l'acquisition d'équipements, elles étudient le sens liant l'ensemble des comportements avec les attitudes exprimant les tensions entre le ménage et son modèle culturel.

Des outils quantitatifs ont été par ailleurs mobilisés pour classifier des « modes de vies résidentiels » à partir de données croisées d'équipement de comportements, de consommation (…) sans toutefois fait explicitement référence à un cadre théorique culturel (voir p. 32 sur les travaux de classification de données).

1.4 Approches multidisciplinaires : apports et perspectives

Les approches spatiales : le lien entre formes urbaines, pratiques et consommations d'énergie

Les cadres présentés ci-dessus ont en commun de situer les caractéristiques du logement (son type, sa taille, les modes d'énergie) comme des facteurs déterminants des pratiques. Les études urbaines (*Urban Studies*) abordent le logement au sein d'échelles spatiales plus larges. Un regard rapide sur ce champ permet d'observer la spatialité des comportements et des consommations d'énergie. Dans un travail de modélisation empirique à l'échelle de la région Île-de-France, Bourgeois et *al*. (2017) croisent les CED avec les facteurs démographiques et les dynamiques résidentielles. En croisant trois dimensions comportementales (niveau d'équipement, intensité d'usage et comportements de régulation) avec des données spatiales des habitats (logements, zone urbaines), ils montrent une forte corrélation entre pratiques domestiques et les dynamiques territoriales (Figure 11).

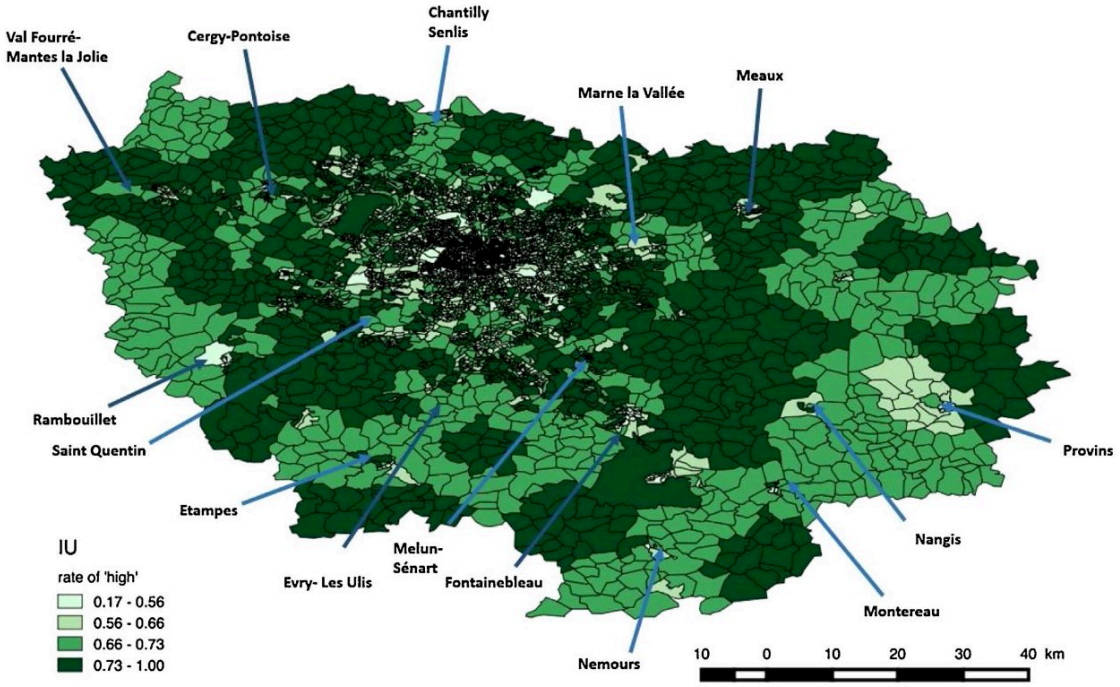


Figure 11 : Simulation de l'indicateur d'intensité d'usage des équipements énergivores (IU) dans la région Île de France. Le modèle est entraîné à partir de données d'enquête de 2010. Source : (Bourgeois, 2017).

Le modèle de Kowsari : illustrer l'interaction forte dans chaque logement entre des phénomènes sociaux, psychologiques, physiques.

Dans son article de 2011, Kowsari fait le constat d'un écart important entre les modèles de comportement et la réalité observée. Il souligne également un intérêt des études économétriques pour mesurer l'effet de variables tel le genre ou l'âge sur la CED, mais il considère que l'impact de ces variables n'indiquent pas ou de manière imprécise la façon dont elles agissent sur les situations. Aussi, Kowsari souligne le contraste entre un intérêt des études quantitatives pour l'étude des effets de variables isolées (le prix, la technologie, l'énergie) alors que la plupart des études qualitatives soulignent leur forte interrelation. Il propose dans son article un cadre de modélisation permettant de caractériser la CED à l'aide de trois dimensions : les « services énergétiques » ; les « systèmes énergétiques » ; les « énergies disponibles dans le logement ». L'innovation de cette contribution est de proposer une modélisation multi-disciplinaire et multi-échelle (individu, ménage-logement, société) permettant d'intégrer les éléments de connaissance issus de sciences humaines tout en ouvrant la voie à une évaluation empirique quantitative.

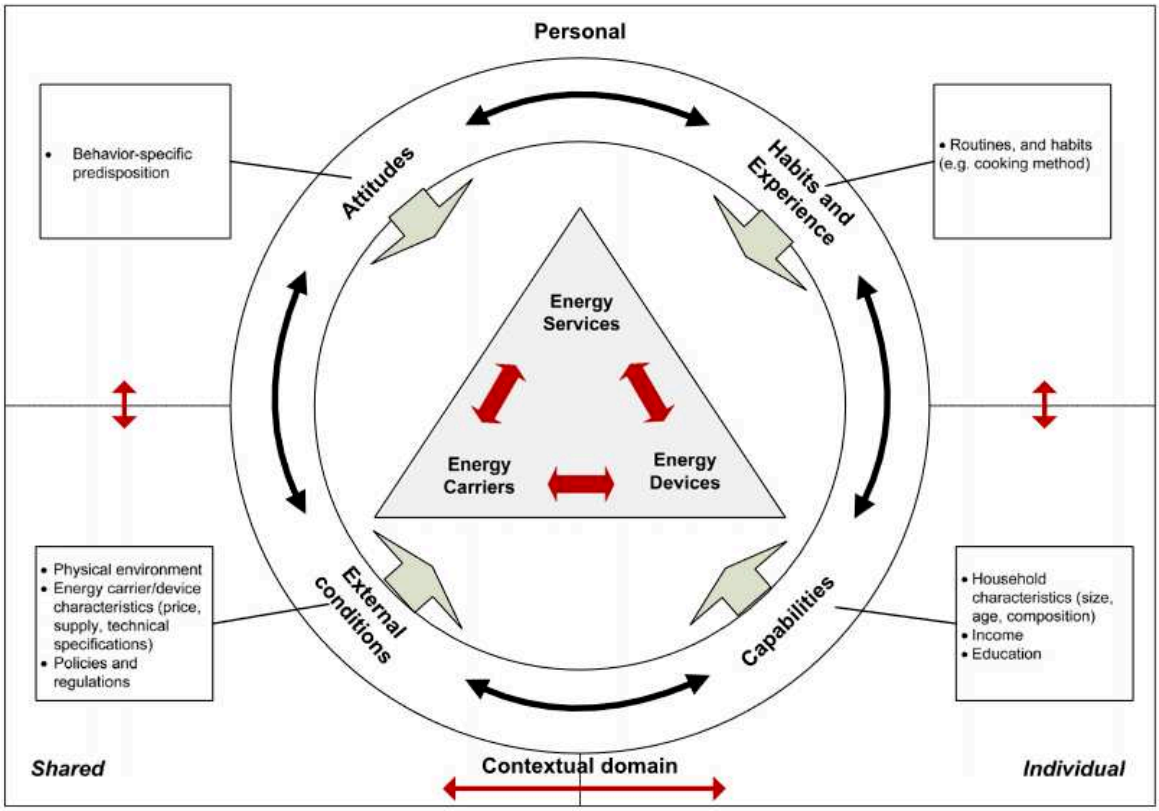


Figure 12 : Cadre de modélisation, de la CED par (Kowsari, 2011). La consommation d'énergie peut être considérée comme le produit d'un système cohérent de services énergétiques, de consommateurs d'énergie et de modes d'énergie disponibles. Ces trois facteurs sont eux-mêmes déterminés par des effets de contexte, individuels et collectifs. La légende originale précise que les variables citées dans le schéma ne sont pas exhaustives mais ont vocation à illustrer le cadre de modélisation proposé. Source : Kowsari (2011).

2. Analyse critique de l'état de l'art

Les différentes familles de travaux qui ont porté sur la modélisation illustrent les différents concepts, la diversité des objectifs poursuivis (compréhension des phénomènes, aide à la décision d'acteurs publics, privés ou même au sein de systèmes automatisés). Nous proposons ici une vision transversale de ces approches afin d'illustrer leurs points de clivage et leurs enjeux heuristiques.

2.1 Un manque de modélisation quantitative intégrant la dimension sociale des consommations d'énergie

2.1.1 Les grandes familles de modèles explicatifs

La revue de littérature des modèles explicatifs permet d'identifier quatre grandes familles présentées dans le schéma ci-dessous (Figure 13). Chacune d'entre elles étudie l'influence d'une catégorie de variables (les systèmes techniques, la psychologie de l'individu, les pratiques sociales, etc.) sur la consommation d'énergie domestique. Parmi elles, les approches anthropologiques et sociologiques ont la particularité de produire des concepts et des théories (appropriation des logements, contextes

résidentiel, définition du confort, parcours résidentiel, pratiques sociales) afin d'illustrer et comprendre les dynamiques des comportements consommateurs d'énergie domestiques.

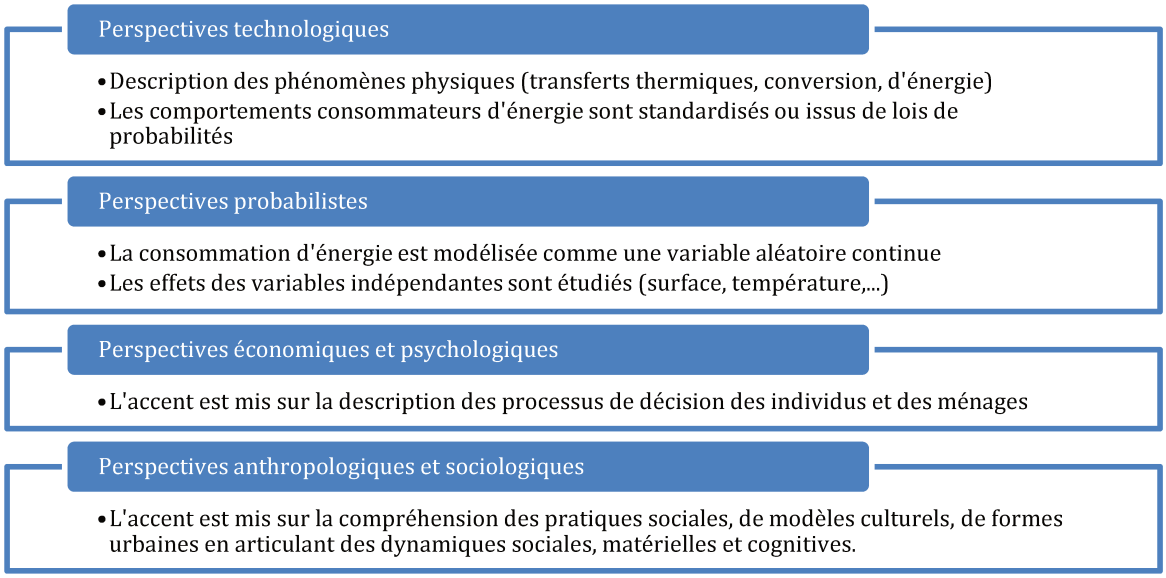


Figure 13 : Synthèse des principales familles de travaux identifiés dans la littérature et proposant un modèle explicatif de la consommation d'énergie domestique.

Nous avons cependant remarqué que les travaux en anthropologie et en sociologie fournissaient des concepts et des théories intéressantes (appropriation des logements, contextes résidentiel, définition du confort, parcours résidentiel, pratiques sociales) pour illustrer et comprendre les dynamiques des comportements consommateurs d'énergie dans les logements.

2.1.2 Le besoin de théorisation des processus de consommation d'énergie domestique

Un premier constat est que les travaux de modélisation de la CED, essentiellement dans le domaine de l'ingénierie ou de la statistique, soit considèrent le choix des individus comme rationnel, soit ne justifient pas ou peu la mobilisation des variables sélectionnées. A première vue, la production massive de données et d'algorithmes (le « big data ») permettant d'appréhender ces comportements pourrait laisser espérer que les analyses empiriques valident ces modèles. Toutefois, face à la réalité empirique, ils font l'objet de nombreuses critiques.

En premier lieu, nous pouvons avancer que toute approche modélisatrice quantitative repose implicitement sur une vision hypothético-déductive du réel. En effet, ce type de modèles repose sur une ensemble complexe de tâches de mise en variable et de quantification et de création de conventions d'équivalence entre réalité et monde mathématique abstrait qui autorise les opérations logiques et les analyses que réalisent les modélisateurs en tant que tels (Desrosières 2001). La sélection des objets du réel et leur « mise en variable » est déjà, de fait, une étape fondamentale du travail de modélisation. On

peut reprendre ici les mots de Gaston Bachelard qui voient dans les instruments de mesure (ici les enquêtes, les capteurs) une « théorie réifiée » (Bachelard 1934).

Ensuite, une autre limite de ce type d'approche est que la construction de modèle prédictifs, à l'aide de critères d'adéquation avec les données (i.e. de précision), ne permet pas en soi d'élaborer une structure mathématique traduisant les processus réels. Il identifie plutôt les cadres heuristiques permettant de réaliser une estimation la plus précise possible du réel. Par exemple, les travaux de (Namazkhan, Albers,et Steg 2020) illustrent la façon dont la construction d'un modèle d'arbre décisionnel par apprentissage supervisé pour estimer la consommation de gaz en Hollande permet d'identifier un ensemble de variables et de règles reposant sur des seuils. Dans ce travail, les auteurs identifient notamment « la taille de la maison, l'âge du bâtiment et le type de résidence (caractéristiques du bâtiment), le revenu du ménage et la situation professionnelle (données sociodémographiques), et plus particulièrement les valeurs égoïstes, les valeurs hédoniques, la conscience environnementale, la perception de la responsabilité environnementale du fournisseur d'énergie et la norme sociale (facteurs psychologiques) » (Figure 14). A partir de ce graphe, les auteurs expliquent cependant qu'« Il est intéressant de noter que le réglage de la température ambiante, en tant qu'indicateur du comportement des ménages en matière de consommation de gaz, n'est pas lié de manière significative à la consommation de gaz des ménages ». Dans leur analyse, les auteurs mettent également en évidence les limites de leur démarche. Ils illustrent ainsi la tension entre l'existence avérée du lien entre le réglage de la température et la CED avec l'absence de données statistiques permettant de valider le modèle prédictif. Une discussion intéressante pourrait être menée pour étudier la qualité de cette variable, son adéquation avec le comportement « réel » des ménages ou encore la corrélation de cette variable avec les celles sélectionnées dans le modèle. Cet exemple permet ainsi de souligner le fait que les corrélations entre les variables ne permettent pas aux modèles de rendre compte de la réalité des processus décrits.

Enfin, une dernière remarque, plus opérationnelle, est de rappeler que les modèles construits sont aussi des objets techniques, insérés dans des espaces sociaux, répondants à une demande. La sélection des variables, des algorithmes, et des théories dépend de cette demande (Varenne et Silberstein 2013). Celle-ci peut être d'identifier des cibles afin d'agir sur les politiques publiques, si le modèle est construit dans un cadre de recherche opérationnelle, ou de participer à la construction d'un argumentaire statistique permettant de la soutenir (Desrosières 2013). Dans ces cadres, les variables sélectionnées reprennent le vocabulaire, les concepts ou *a minima* sont compatibles et accessibles avec le contexte opérationnel dans lequel ils sont utilisés. Dans le cas où les modèles de CED sont utilisés pour soutenir une décision de rénovation sur un logement, les variables doivent décrire *a minima* la dimension physique et fournir une estimation la plus transparente pour les parties prenantes de la décision. Toutefois, dans une perspective heuristique, le choix des variables et de la stratégie de modélisation est plus libre, mais aussi conditionné par l'ancrage théorique et disciplinaire car les résultats et les méthodes sont évaluées par les pairs.

Mais la revue de la littérature a également été l'occasion de questionner nos propres pratiques de modélisation et de souligner la nécessité d'expliciter les bases de l'approche théorique déterminant les modèles que nous avons construits (Wise et Shaffer 2015).

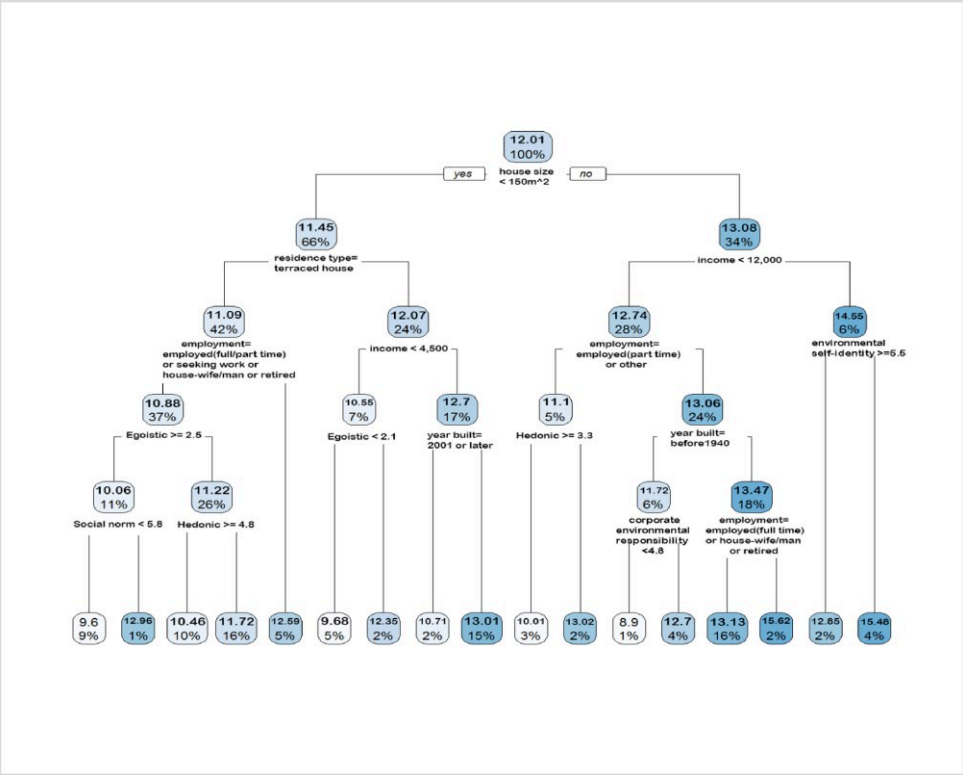


Figure 14 : Image tirée de l'article de (Namazkhan et al., 2020) avec la légende suivante (traduite) : « Arbre de décision pour expliquer la consommation totale de gaz des ménages néerlandais. Le premier chiffre de chaque case de nœud représente la consommation moyenne de gaz des ménages dans cette branche, tandis que le deuxième chiffre de chaque case de nœud indique le pourcentage de ménages de l'échantillon qui se retrouvent dans cette branche (...). »

2.1.3 L'intérêt d'une approche par les contextes résidentiels

L'examen de la littérature montre que les prévisions robustes de la CED des logements restent à produire. Les modèles estimant la consommation de parcs de logements parviennent à estimer la CED annuelle de chauffage de quelques centaines de logements représentatifs avec une erreur moyenne de l'ordre de 10% (Allibe, 2012). Toutefois, ces mêmes modèles contiennent des erreurs d'estimation d'un facteur 2 ou 3 à l'échelle d'un logement ou d'un bâtiment. A l'évidence, ils sont aujourd'hui peu performants à l'échelle du logement et « les publications sont discordantes sur les modèles qui ont une meilleure capacité prédictive, ce qui confirme que la capacité d'un modèle à prédire la réalité dépend fortement à la fois de l'étude de cas considérée et des indicateurs de performance utilisé » (Gaetani, Hoes,et Hensen 2016). Il faut rappeler cependant que la performance des modèles construits est de l'ordre de l'estimation et non de la prédiction. Les modèles ont été développés sur la base d'un sous-ensemble de l'enquête des ménages et des logements qui n'a pas été utilisé pour le calcul de leurs paramètres. Leurs performances prédictives (c'est-à-dire sur des données issues d'autres enquêtes, sur

d'autres espaces géographiques et temporels) sont rarement évaluées. Elles permettraient pourtant de faire apparaître la dépendance temporelle et spatiale des estimateurs construits.
Cependant, si la validation empirique de modèles à l'échelle agrégée fournit de bons estimateurs et identifie, pour les opérateurs, des cibles pour l'action (en l'occurrence la rénovation du parc de bâtiments et le remplacement des systèmes de chauffage peu efficaces, la limitation de grandes surfaces et particulièrement la préférence pour le logement collectif), ils ignorent pour autant la portée explicative des processus de consommation. En effet, ils fournissent des estimations des effets moyens, mais ceux-ci dépendent *a priori* de l'échantillon considéré. Ainsi, les recherches actuelles étudiant le gap de performance énergétique (Energy Performance Gap - EPG) et celles cherchant à segmenter et construire des cibles pour l'action publique (des sous-ensembles de ménages, de logements) illustrent la difficulté à dégager une connaissance unifiée de la CED. Par ailleurs, les travaux qualitatifs et quantitatifs sur les comportements domestiques soulignent le caractère multi-échelle du phénomène de la CED (dépendance spatiale et temporelle liée à des différences, par exemple culturelles, climatiques, etc..). Dans notre approche, les comportements domestiques, les pratiques résidentielles, les normes sociales et culturelles, les infrastructures techniques d'accès à l'énergie et les prix sont des variables sociales macroscopiques qu'il est nécessaire d'étudier comme telles, elles sont à la fois « structurantes et structurées » (Bourdieu, 1979) dans le processus local et social de la CED. Les recherches présentées soulignent également la dimension systémique des stratégies comportementales des ménages. Ceux-ci tentent de s'approprier leur logement et les systèmes pour parvenir à déployer (ou non) leur perception du confort définie à travers la décoration, la présence, la réalisation d'activités, l'usage plus ou moins intensif d'un ensemble d'équipements, et des pratiques de régulation énergétique. Cette architecture de comportements décrite par Bonnin et Subrémon peut être comprise comme un élément observable d'une interaction entre le matériel et le social.

Les différents modèles explicatifs mobilisent des facteurs physiques, socioculturels, psychologiques et économiques pour expliquer les comportements et les consommations d'énergie domestique. En prenant le parti d'aborder les consommations d'énergie sous un angle social, c'est-à-dire en observant les comportements dans l'espace domestique en tant que pratiques sociales, il devient possible de faire apparaître le contexte résidentiel dans lequel elles s'exercent.

Le contexte résidentiel est un concept issu des études urbaines, notamment celles portant sur la mobilité et les pratiques résidentielles au sens large. Pour Lévy (Lévy 1998), le contexte résidentiel est l'ensemble des « caractéristiques qui définissent la position socio-démographique et socio-économique du ménage et les caractéristiques physiques du logement ». Cette définition permet de connaître la position géographique du logement dans un espace résidentiel, la position sociale de l'habitat grâce aux caractéristiques du ménage et du logement et de mener une réflexion sur les pratiques de mobilité résidentielle relativement à ces contextes. Dans cet exemple les variables utiles sont : le statut du logement (public/privé), le nombre de pièces et le niveau de confort, le type d'immeuble et la date de

construction, l'âge de la personne de référence du ménage, sa catégorie socio-professionnelle, son statut d'occupation, sa nationalité, son statut d'activité, et la taille du ménage.

Dans la littérature adoptant une perspective « habitante » pour expliquer les consommations d'énergie, le contexte résidentiel est aussi convoqué, mais avec une définition différente de la précédente. Ici, il s'agit à la fois de caractériser l'espace physique où se déroulent des pratiques sociales relevant d'un modèle culturel, et l'espace social hiérarchisé. Dans les travaux de Bonnin et Subrémon, les caractéristiques de ce contexte se déclinent selon des variables définissant l'adaptation du logement aux besoins du ménage : surface, nombre de pièces du logement et par personne, type de logement, statut d'occupation du logement (secteur privé ou public, propriétaire ou locataire), zone urbaine, type de chauffage et mode d'énergie principal, niveau d'isolation thermique. Et des variables caractérisant le ménage : âge de la personne de référence (PR), revenus du ménage, statut d'activité de la PR, catégorie socio-professionnelle (CSP) de la PR, nombre d'enfants, composition du ménage.

2.2 Les enjeux théoriques et techniques d'une modélisation intégrant les pratiques sociales

2.2.1 La difficulté de réconcilier les approches individualistes et structuralistes

Les modèles qui intègrent les pratiques domestiques peuvent être regroupés en trois types de familles : (1) ceux qui valorisent des scénarios standardisés ou probabilistes qui ont plutôt vocation à introduire un effet plus ou moins cohérent avec les réalités empiriques (les scénarios standardisés sont une approximation grossière des scénarios d'occupation et d'activité domestique (Widén, Molin,et Ellegård 2012), (2) ceux qui valorisent des formulations à l'échelle individuelles (utilitaristes, modèles de rationalité, influence du contexte, etc) ; et (3) ceux s'inspirant d'une compréhension sociologique des comportements domestiques. Cette différenciation se retrouve, d'une part, dans les solutions explorées pour tenter d'agir sur les CED et, d'autre part, dans les propositions basées sur une conception individuelle de la CED (Hampton et Adams 2018). Les recherches basées sur des petits échantillons de ménages montrent toutefois que la rationalité et le calcul utilitaire sont difficiles à transcrire dans les activités quotidiennes puisque les recherches portent sur une déclinaison contextualisée d'une définition du confort, influencé par des normes sociales et des contraintes physiques, temporelles et financières. Malgré l'intérêt des modélisations en psychologie pour étudier les contextes locaux, elles ne seront pas prises en compte du fait d'un manque de travaux empiriques visant à comprendre la diffusion et la structure des comportements domestiques, et leur influence sur la CED.

2.2.2 La reconnaissance de la dimension symbolique des systèmes et des contextes matériels dans la construction des pratiques et des consommations d'énergie

La revue de littérature a permis d'observer que la construction des modèles explicatifs de la CED repose principalement sur deux paradigmes scientifiques. D'une part les approches quantitatives, sont fondées sur la construction de modèles basés sur des critères d'adéquation avec un ensemble de données collectées. D'autres part, les approches compréhensives considèrent les comportements et la CED comme des observables qui doivent être traduits et réinterprétés pour expliciter les processus à l'œuvre. Dans ces derniers travaux de recherche, c'est la rigueur dans la collecte de données ciblées, leur analyse et leur structuration dans un cadre théorique défini qui détermine la qualité d'un modèle. Si les deux approches ont fourni des résultats significatifs, il est frappant d'observer qu'ils mobilisent des concepts et des théories distincts (« appropriation » et « modèles culturels » pour les approches compréhensives versus « élasticité de la demande » pour des travaux positivistes par exemple). Par ailleurs, les recherches elles-mêmes portent sur des échelles spatiales très différentes (réduites pour les premières, et plus larges pour les secondes). Cependant, le croisement de ces deux approches dans un modèle commun reste, à ce jour, à explorer et à construire et, ceci, quelle que soit l'échelle d'observation.

2.2.3 Qualité, rareté des données et difficultés méthodologique de la quantification des pratiques sociales

Un autre point notable est que la diffusion des données produites dans le domaine résidentiel est particulièrement encadrée (afin d'éviter la localisation des minorités ethniques). Elles sont rares et bien protégées ce qui implique la nécessité de produire des enquêtes ad hoc dès lors que l'on désire travailler sur des périmètres et des échelles déterminés. Par ailleurs, le domaine résidentiel recouvre le champ de nombreuses disciplines. Les données constituées relèvent des problématiques propres chacune d'entre-elles et sont difficilement transposables (Desrosières, 2001). Par exemple, si la plupart des enquêtes portent sur le coût de l'énergie, les comportements environnementaux et de régulation (approches économiques et psychologiques), elles comportent peu de questions sur la définition du confort des ménages ou l'appropriation du logement (approches culturelles). Par ailleurs, ces travaux sont très souvent ponctuels et ne sont pas réalisés dans une perspective longitudinale (c'est à dire le suivi des mêmes ménages et logements à des échéances temporelles régulières), en ne permettant pas de mettre en évidence les évolutions comportementales des ménages face aux transformations de l'immeuble ou du logement (cas d'une rénovation par exemple).

La qualité des données est également un point peu abordé dans les études citées (Kavgic et *al.*, 2010) Quelques études apportent néanmoins des éléments quantitatifs pour apprécier de l'amplitude des incertitudes, de l'impact des choix méthodologiques de réduction (spatiale et temporelle) des comportements domestiques, de leur influence sur la précision des modèles d'estimation, le poids statistique relatif des variables. Sur la qualité des données d'énergie collectées par enquête, une étude

allemande avait, en 2005, estimé que l'erreur de collecte sur les factures énergétiques des ménages était faible en ce qui concerne le gaz et l'électricité, mais plus importante pour le bois, le fioul ou le gaz liquéfié ainsi que les énergies renouvelables (Christiansen et *al.*, 2005). De leur côté, (Warriner, McDougall et Claxton, 1984) ont estimé que les erreurs dans les données d'enquête se situaient entre 10 et 30% de la consommation réelle selon le type d'énergie. D'autres études ont montré que des biais cognitifs (désirabilité sociale, manque de connaissance, etc.) pouvaient aussi agir sur la construction des réponses et provoquer des écarts avec la réalité. Enfin, à l'échelle d'un quartier résidentiel allemand, (Nouvel et *al.*, 2017) montrent que la fonction du bâtiment, l'année de construction, l'état de rénovation et le type de résidence sont des variables indispensables pour prévoir la CED avec une erreur d'estimation de l'énergie annuelle du quartier de moins de 30% (toutes les autres variables du modèle étant supposées connues).

Un autre problème lié aux données réside dans la diversité des méthodes de quantification de la CED. Dans la littérature, elle est « mise en variable » tantôt par domaine de consommation (chauffage, ventilation, éclairage, ECS, loisirs, etc.), tantôt par type d'énergie (électricité, gaz, bois, etc.), en énergie primaire ou en énergie finale, voire en unité monétaires (euros, dollars). Aussi, la CED totale peut être considérée comme un tout en consommation annuelle (en kilowattheure, kWh), ou bien être normalisée par mètre carré de surface habitable (kWh/m²). Elle peut être également calculée en fonction du nombre de personnes du ménage (kWh/personne) ou par unité de consommation (kWh/UC). Dans la littérature, peu d'articles s'intéressent à la question de la métrique qui, pourtant, semble conditionner largement les décisions publiques. Dans un travail de modélisation entrepris en Suède (von Platten, Mangold,et Mjörnell 2020) montrent que « qu'en mesurant la consommation d'énergie par habitant plutôt que la consommation d'énergie normalisée par zone, les bâtiments inefficaces sur le plan énergétique se trouvaient dans les centres-villes à revenu élevé plutôt que dans les banlieues à faible revenu des villes suédoises » (traduction de l'anglais par l'auteur). En s'appuyant sur un travail de modélisation statistique à partir de données sur la France, Lévy et *al.* (2018b) montrent, quant à eux, une variation selon que la CED d'un même ménage soit calculée par m² ou par personne. Ils soulignent notamment qu'« un ménage occupant un logement de moins de 70 m² a, toutes choses égales par ailleurs, 2,5 fois moins de chances d'être un faible consommateur par m² qu'un ménage vivant dans un logement de 100 m² ou plus, et 1,2 fois plus de chances d'être un faible consommateur par personne ». Finalement, il apparaît que l'intégration simultanée des trois métriques (kWh, kWh/m², kWh/personne) permettrait d'améliorer le diagnostic énergétique. La lecture de ces indicateurs est en effet plurielle : au-delà des quantités d'énergie entendues au sens physique, ils rendent compte du volume de « services énergétiques consommés » (kWh), de l'intensité d'usage énergétique du logement par le ménage occupant (kWh/m²) et de l'intensité d'usage par les individus composant le ménage, dans le logement (kWh/p).

2.3 Le développement « récent » des approches intégratrices

2.3.1 Une approche des consommations par le « système énergétique domestique »

Les contributions disciplinaires présentées ci-dessus ont permis d'identifier des phénomènes physiques, sociaux, économiques, psychologiques déterminants. Les tentatives d'intégration des perspectives disciplinaires dans un cadre de modélisation commun ne sont pas récentes comme en témoigne le modèle de Van Raaij publié en 1983. Toutefois ce champ est renouvelé depuis une quinzaine d'années autour de travaux articulant des concepts de « mode de vie » ou « culture ». Ainsi, Janet Stephenson et *al.* (2010) expriment leur surprise sur le fait que les travaux de Loren Lutzenhiser sur l'identification de modèles culturels et la caractérisation des gestes domestiques des ménages (Lutzenhiser, 1992) n'aient pas été poursuivis. Ils soulignent néanmoins que ces premiers travaux n'étaient pas assez formalisés pour être considérés comme une théorie : il aurait fallu alors identifier un ensemble de concepts et d'interrelations pour permettre des réplications et des tests empiriques. On remarquera d'ailleurs que les travaux de psychologie sociale intègrent la dimension sociale des comportements énergétiques en tenant compte du « coût » ou de la « contrainte » sociale, liée au poids (ou non) des normes sociales pesant sur les individus. Cependant, ces modèles contextualisés ne traduisent pas l'évolution de ces pratiques (i.e. leur émergence, leur reproduction, leur disparation, leurs mutations). En revanche, les approches « culturelles » s'appuient sur l'étude croisée de trois dimensions pour caractériser une culture énergétique d'un ménage : comportementale (les comportements énergétiques et non énergétiques déployés dans le logement) ; matérielle (la qualité physique du logement, l'ensemble des équipements situés dans le logement) ; cognitive (les connaissances, les croyances et les normes). Le travail de Subrémon et Bonnin montre comment cette approche tridimensionnelle de variables met en évidence l'existence d'un « système énergétique domestique » (SED). Dans les faits, peu de travaux abordent la question des consommations énergétiques domestiques comme un système, mis à part peut-être la synthèse réalisée par Kowsari (2011) évoquée plus haut.

2.3.2 Les approches par classification de données : entre intérêt scientifique et enjeux méthodologiques et techniques

L'intégration d'approches économiques, psychologiques, d'ingénierie conduit dans les faits à appréhender des dynamiques croisées. C'est le cas de travaux portant sur la construction de « cas d'études » voire d'« archétypes » (entendus comme des « prototypes des réalités visibles du monde », selon le Larousse), construits par segmentation de tables de données. Cette méthode associe des phénomènes difficiles à modéliser de manière analytique (physiques, psychologiques, sociaux…). Mais elle s'appuie également sur des concepts communs à la fois disciplinaires et pluridisciplinaires dans un échange entre ces deux approches. Toutefois, les travaux de classification souffrent de plusieurs limites techniques et théoriques.

En ce qui concerne les limites techniques, la classification nécessite de traiter à la fois des données qualitatives et quantitatives. On évoquera alors la construction d'un corpus de données « mixtes » (« mixed data »). Par ailleurs, une part croissante de travaux croise des données issues d'enquêtes, d'observations techniques et comportementales, de mesures issues de capteurs ou de compteurs, dont la temporalité et la spatialité sont différentes. L'intégration de ces données au sein d'un même cadre de modélisation est un enjeu majeur pour ces travaux. Pour contourner cette difficulté, de nombreux auteurs procèdent par une catégorisation des variables quantitatives, ce qui engendre un biais important dans la méthodologie. C'est la raison pour laquelle nombre des recherches recensées dans cet état des lieux mobilisent une « analyse tandem » qui associe une analyse factorielle (voir annexe pour plus de détail) pour construire un espace synthétique de composantes principales orthogonales, et une typification des ménages par Classification Ascendante Hiérarchique (CAH). Toutefois, les méthodes d'analyse factorielle et de CAH ne sont pas nécessairement associées. L'une et l'autre peuvent être utilisées indépendamment. Ce qui peut introduire des biais dans les méthodes de l'analyse tandem. Ainsi, de Soete et Carroll (De Soete et Carroll, 1994) notent que l'espace des composantes principales orthogonales ne permet pas toujours la représentation de l'information taxonomique.

Concernant les limites théoriques, la littérature en sociologie de l'énergie a produit plusieurs typologies de comportements basées sur des enquêtes ethnographiques. Les travaux de modélisation quantitative ont, quant à eux, produit plusieurs typologies de comportement, mais il est difficile de faire le lien entre les classes produites et les contextes résidentiels.

Pour conclure cet état de l'art, il nous semble important de souligner que ces différentes approches modélisatrices, quantitatives et qualitatives, convergent vers quelques questions de recherche :

- Dans quels contextes résidentiels s'exercent les pratiques sociales ?
- Dans quelle mesure existe-t-il un lien entre ces pratiques ?
- Dans quelle mesure les contextes résidentiels permettent-ils d'expliquer les pratiques énergétiques ?
- Dans quelle mesure les contextes résidentiels permettent-ils de comprendre les consommations d'énergie ?

- Ces questions de recherche posent la question de l'articulation entre comportements, abordés comme des pratiques sociales, et les caractéristiques des ménages et des logements. En d'autres termes, **dans quelle mesure les « contextes résidentiels sociaux et matériels » permettent-ils d'expliquer les niveaux de consommation domestique d'énergie finale ?**

3. Proposition de recherche : une modélisation des consommations d'énergie basée sur les contextes résidentiels

> *" The dwelling is actually a large artefact into which groups are fit. In its form, it comes to reflect, the sensibilities of its residents, the material realities of its surroundings and the social structures that both contain it and call it home" (L. Lutzenhiser, 1992)*

Notre problématique place « l'habiter » comme un déterminant central des pratiques domestiques. Par extension, c'est toute la consommation d'énergie qui est concernée, en tenant compte des contraintes physiques (la surface, l'isolation thermique des murs, l'efficacité des équipements, leur type et leur nombre) et sociales (normes).

Il ne s'agit pas de réaliser un exercice de modélisation de « physique sociale » où l'individu, les ménages et leurs comportements seraient réduits à quelques variables. Il ne s'agit pas non plus d'affirmer que l'individu et les ménages seraient isolés de tout contexte matériel ou social et que leurs actions ne dépendraient que de variables résumant les individus et leurs contextes. L'état de l'art plaide davantage pour la nécessité de construire des modèles quantitatifs permettant de caractériser et de décrire les influences matérielles et sociales qui s'exercent sur les ménages et leurs pratiques des lieux (Stock 2003).

3.1 Les hypothèses de modélisation de la CED

Nous proposons dans cette thèse de mettre à l'épreuve le cadre de modélisation interdisciplinaire proposé par Kowsari (2011). Celui-ci transcrit, dans un périmètre adapté à la modélisation, à la fois la dimension matérielle des modèles d'ingénierie et les dimensions habitantes et culturelles des approches sociologiques de la CED. Nous nous proposons d'adapter ce modèle à partir d'hypothèses simplificatrices permettant, en partie, de le valider empiriquement.

Hypothèses

- La **CED** est un phénomène observable dans des **logements** occupés par des **ménages**. Par convention, la relation du couple ménage-logement est dénommée « **situation d'habitation** », décrite par l'ensemble des variables servant à l'explication des processus de consommation d'énergie identifiées dans la revue de littérature.
- La CED est issue de **comportements consommateurs d'énergie** des **individus** composant le **ménage** ou éventuellement par le biais **d'équipements** autonomes. Le volume de la CED dépend de la performance énergétique des équipements utilisés.
- Ces **comportements consommateurs d'énergie** sont des pratiques effectuées lors de séquences **d'activités** (se nourrir, laver le linge, se détendre, etc.)
- Ces **activités** découlent d'une perception du **confort** propre à chacun des **ménages**. Elles sont contraintes par **l'environnement** matériel, financier et culturel. Elles peuvent également comporter des pratiques non-consommatrices d'énergie.
- Le **confort** relève d'un ensemble de **normes** qui donnent au **ménage** un **sentiment de bien-être**, verbalisable à travers des **attitudes positives ou négatives** (confort thermique, perception d'être à l'aise chez soi, en sécurité etc..). Du point de vue de la modélisation, il s'agit d'une variable latente

locale, caractérisable à partir de l'étude croisée des comportements, des matériels et des discours des habitants.

- Le **confort** est une **norme** contraignante et motrice pour le ménage qui le décline dans son **appropriation** du **logement** : les **comportements consommateurs d'énergie** évoluent dans le temps avec **l'appropriation** progressive du **logement.**
- A travers la réalisation des **activités** domestiques, les **individus** s'inscrivent dans un « habiter » **confortable** tout en participant à la **reproduction de normes** des manières d'habiter (**confort**), transmises à leurs réseaux sociaux (enfants, amis, voisins, invités).
- Le **confort** est ainsi perçu comme un phénomène local (dans sa réalisation concrète) et global (ou sens où il est diffusé). Il n'est pas observable en l'état : c'est une définition latente qui n'est perceptible que par la mise en regard des **comportements** du ménage, de l'**environnement** matériel, de la CED mesurée, et des **attitudes** réflexives du ménage.
- Un même **comportement** peut renvoyer à plusieurs modèles de **confort** perçus dans un contexte précis[20].
- Les **activités comportementales domestiques** peuvent être interprétées comme la réalisation locale de pratiques **sociales**. Toutefois, les **assemblages de comportements** des ménages sont globalement cohérents car ils doivent être réalisés selon des **normes**, et dans un **cadre de ressources** (financières, temporelles, cognitives, matérielles) contraint. L'assemblage des **comportements domestiques** fait donc l'objet d'une **négociation** et d'une appropriation au gré des contraintes et des ressources locales.
- La considération croisée des **comportements domestiques** (**pratiques**), de la dimension cognitive (des savoirs, compétences, croyances) et de la dimension matérielle (équipements, décors, matériels, qualité du bâtiment) permet de définir un « **style de vie résidentiel** », qui renvoie à un **modèle culturel** décliné dans un contexte local.
- La **CED** d'un ménage est ainsi issue de **l'interaction** locale entre des évolutions spatiales et temporelles hétérogènes (individus, ménage, logement, quartier, normes, infrastructures énergétiques). Le terme **d'interaction** définit la correspondance entre le champ de la modélisation et celui de la sociologie de l'habitat où elle désigne le processus **d'appropriation** d'un **logement** par un **ménage**.
- La diversité locale des **interactions** ne peut faire l'objet d'une modélisation explicite à une échelle large, dans la mesure où les processus n'agissent pas tous dans le même cadre spatial et temporel. L'insécurité d'un quartier ou d'un immeuble, les défauts récurrents des systèmes de régulation thermiques, le sentiment de précarité ou d'isolement social sont des phénomènes locaux dont

[20] Pour reprendre un exemple concret tiré du travail de Marguerite Bonnin : l'ouverture des fenêtres en période hivernale peut renvoyer à l'impossibilité pour le ménage de contrôler le système de chauffage (par exemple collectif et défectueux), mais aussi à une pratique d'aération naturelle ancrée dans le modèle culturel du ménage.

l'impact sur la CED est avéré même s'ils n'agissent pas nécessairement en interaction. Le corollaire est que selon l'échelle spatiale et l'espace considérés, les dynamiques observées seront plus ou moins importantes et génèreront des interactions variables selon les **situations d'habitations** étudiées.

- Même si ce cadre de modélisation s'appuie sur la théorie des pratiques, il ne permet pas de rendre explicitement compte de la circulation des usages (par exemple d'hygiène ou de chauffage). Il la médiatise par leur intégration dans des stratégies comportementales locales construites par les ménages dans leur contexte résidentiel.
- Concernant, l'influence du prix de l'énergie, nous rejoignons l'hypothèse faite par (Hackett, 1991) : le prix marginal est une contingence (observable donc) mais le facteur latent expliquant la CED est le **style de vie résidentiel**. Le prix est cependant corrélé avec le style de vie résidentiel et la situation d'habitation.

En résumé, nous proposons de comprendre la CED comme étant le produit d'une interaction locale entre des phénomènes ayant des dynamiques spatiales et temporelles à une échelle plus large. Influencée par les travaux de socio-anthropologie de l'habitat, notre définition de l'interaction regroupe la somme des gestes d'appropriation du logement par le ménage, au regard de ses normes culturelles réalisées dans un contexte de ressources sociales, économiques, matérielles, symboliques et d'un logement situé spatialement et socialement. L'appropriation du logement par le ménage s'inscrit dans la construction d'un système stable de pratiques, de croyances, savoirs, compétences et matériels. Autrement dit, le modèle s'appuie sur une double hypothèse :

(1) Il existe des assemblages de comportements domestiques, d'équipements, d'attitudes, de consommation d'énergie récurrentes et qui définissent des **systèmes énergétiques domestiques.**
(2) Ces modèles culturels sont fortement liés à des **situations d'habitation.**

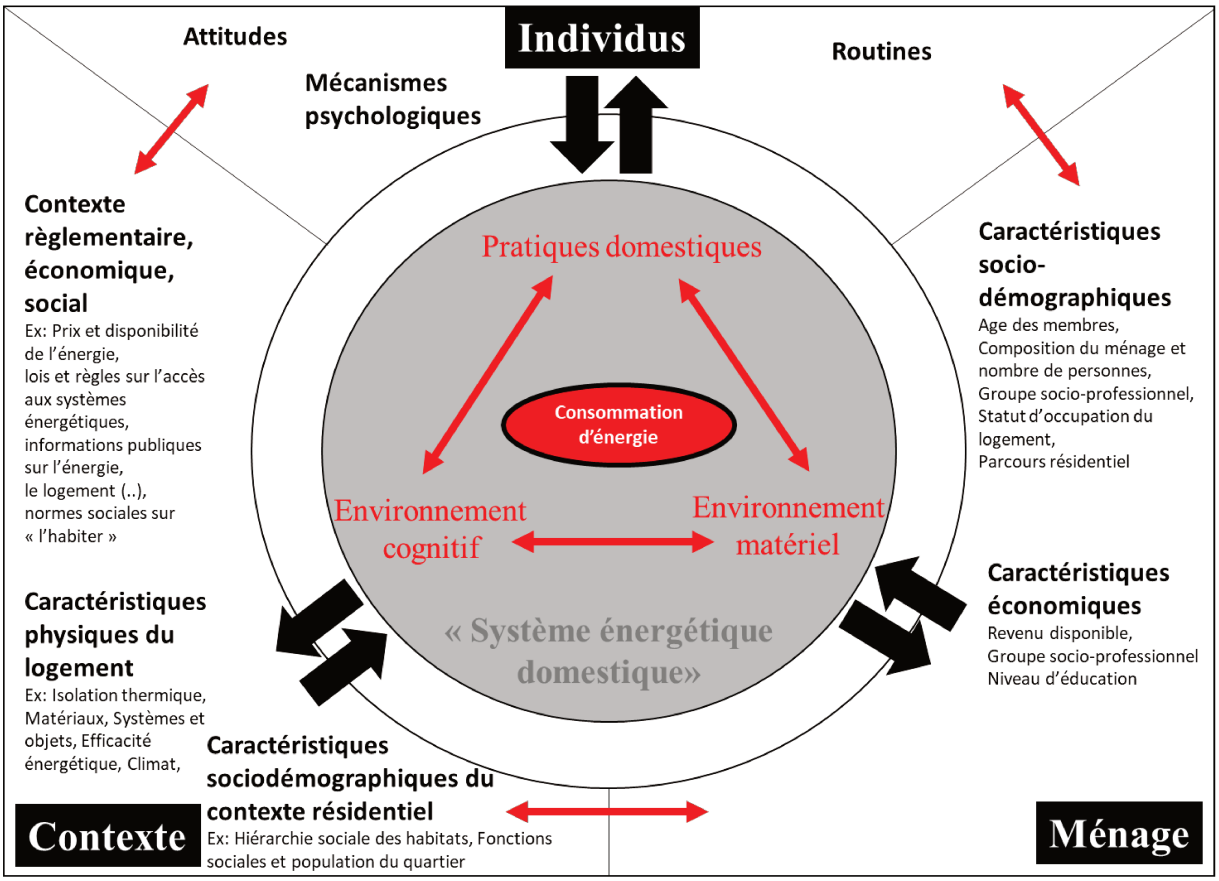


Figure 15 : Adaptation du modèle de Kowsari de la consommation d'énergie domestique. L'interaction entre trois groupes de variables associés au ménage, logement et au contexte permettent de rendre compte d'un style de vie résidentiel qui est caractérisé par trois dimensions : les croyances, savoirs, compétences (cognitif), les comportements réalisés dans l'espace domestique (pratiques), et les équipements, les décors, l'environnement thermique (matériel) dans lequel évolue le ménage. (Source : Auteur, adapté de Kowsari, 2011).

Définition d'une « situation d'habitation »

La thèse repose sur le concept de « contexte résidentiel » qui est inspiré des travaux de sociologie de l'habitat. Les « types résidentiels » construits par exemple par Lévy (Lévy, 1998) décrivent sont des idéaux-types (non institutionnels) de ménages occupant des logements et partageant des styles de vie comparables. Dans ses travaux, Lévy mobilise les types résidentiels pour décrire le contexte de l'habitat et expliquer les dynamiques croisées des évolutions des départements, des communes et des quartiers (et de leur position dans l'espace résidentiel) ainsi que les pratiques de mobilité résidentielle des ménages. Un type résidentiel qualifie la position sociodémographique et socioéconomique du ménage dans le contexte sociogéographique, ainsi que les caractéristiques physiques de l'habitat qu'il occupe.

Dans le prolongement de cette approche théorique, nous proposons de considérer le logement et le ménage dans un système permettant d'appréhender les logiques d'appropriation. Les variables identifiées pour le caractériser sont les suivantes.

Variables caractérisant le logement : la taille du logement (surface, nombre de pièces), le type de logement, le nombre de pièces par personne, le mode de financement de l'immeuble (privé ou public),

le statut d'occupation du logement (en accession ou en location), la zone urbaine, le mode de chauffage et d'énergie principale, le niveau d'isolation thermique

Variables caractérisant le ménage : âge de la personne de référence (PR), revenus du ménage, statut d'activité de la PR, catégorie socio-professionnelle (CSP) de la PR, le nombre d'enfants, situation familiale.

La définition de cet assemblage est aussi liée aux discussions que j'ai eues avec Marina Launay, doctorante en ergonomie au laboratoire CRTD. L'étude de l'activité au sein des espaces domestiques l'a amenée à considérer dans son cadre analytique un espace pour analyser l'activité d'« habiter » et ainsi de proposer la « situation d'habiter » comme un concept dans son travail de thèse. Nos problématiques de travail et les échelles spatiales étaient différentes : il s'agissait pour elle de comprendre l'activité des habitants à une échelle fine, et pour moi de modéliser la consommation d'énergie en intégrant une les pratiques sociales à des fins heuristiques et applicatives et sur des échelles spatiales plus larges. Il me paraît important ici de la remercier, d'expliciter l'influence de nos discussions sur ma proposition de modélisation et aussi d'observer la convergence des approches qui opèrent toutes deux des rapprochements entre les groupes de variables caractérisant les ménages et les logements.

3.2 Présentation du cadre théorique et de la méthodologie

3.2.1 Cadre théorique de la recherche

Genèse et organisation du travail de recherche multidisciplinaire

Par souci de transparence et de rigueur scientifique, et pour réaliser un exercice réflexif, il nous paraît important de restituer les principaux éléments ayant fondé puis guidé cette recherche. Ce regard permet de rendre compte du foisonnement des questionnements individuels, souvent corrélés aux ancrages disciplinaires des travaux mobilisés, de la convergence de leurs questionnements et de leur méthodologie. A l'origine, mon souhait était de réaliser un travail de thèse interdisciplinaire sur des aspects énergétiques. La lecture de travaux de recherche en anthropologie, en sociologie de l'énergie, en modélisation statistique m'a permis de construire un cadre conceptuel transversal, des outils (de classification), et une démarche de validation empirique, sur la problématique de la meilleure connaissance de la CED (Figure 16).

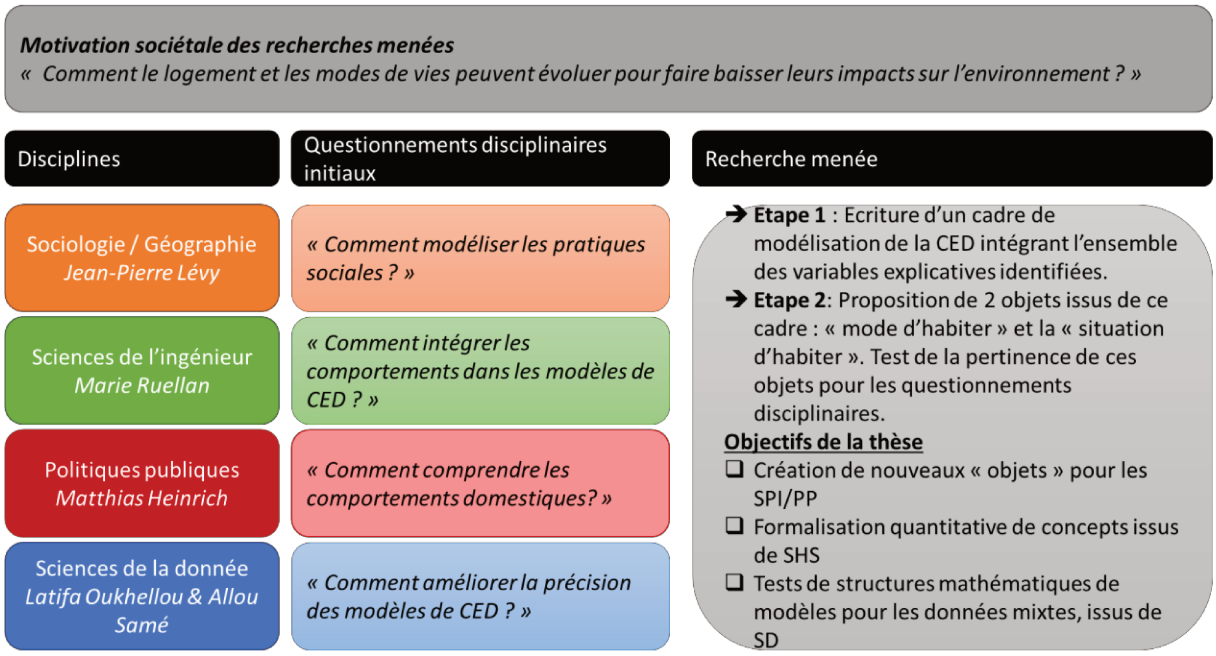


Figure 16 : Schématisation de l'organisation de l'équipe ayant contribué à ce travail de recherche, et des questionnements disciplinaires et individuels posés en amont de ce travail. Source : Auteur.

Regard critique sur la réalisation d'un travail de modélisation quantitative en sciences sociales

> *Just like some politicians will say, "Work toward peace, prepare for war," I will say, "Work toward models, but prepare for description." (Martin 2018)*

La critique de la pratique de la modélisation quantitative est courant en sciences humaines car elle soulève plusieurs difficultés.

Une difficulté ontologique : les variables construites et mobilisées en recherche quantitative sont supposées « représenter » le réel. Toutefois, comme le rappelle Alain Desrosières (2001) le réel ne se donne pas à travers les variables des enquêtes, mais il est construit, négocié, réactualisé au gré des objectifs des acteurs impliqués dans le processus de quantification. Ensuite, la recherche en sciences humaines montre que les variables ont une valeur localisée dans le temps et l'espace et que leur diversité, leur plasticité et la complexité des interactions qui les lient empêchent toute formalisation, rigidification dans un ensemble de variables quantitatives sans perdre la finesse de la compréhension des processus. En ce sens, les approches quantitatives se voient reprochées par les tenants des approches qualitatives une « distance » trop importante avec le réel pour pouvoir produire une connaissance scientifique valide.

> *« La statistique n'explique rien – mais elle fournit des éléments potentiels d'explication » Lebart et al., 1995, p. 209.*

Deux difficultés épistémologiques : la recherche de causalité et le critère de preuve. Pour les recherches quantitatives « traditionnelles », l'articulation des variables au sein de modèles mathématiques (qu'il s'agisse de traitements statistiques élémentaires ou de systèmes d'équations plus avancés) permet de mettre au jour des relations de causes à effets. La preuve de la causalité, au cœur de l'activité scientifique

est toutefois très variable selon les disciplines (biologie, physique, sociologie). En particulier, dans une recherche quantitative la véracité des résultats est associée à une bonne performance empirique : le modèle « reproduit » correctement le réel mesuré (on mesure cette précision à l'aide de critère agrégés, comme l'erreur moyenne absolue ou l'erreur quadratique moyenne). En recherche qualitative, c'est plutôt la rigueur scientifique de la recherche qui est gage de la qualité des résultats produits (Gohier, 2004). Des critiques majeures existent sur ces deux méthodes de démonstration de la preuve. Les modèles quantitatifs peuvent en effet fournir de bonnes performances d'estimation ou de prédiction sans pour autant restituer une explication satisfaisante de la réalité (les corrélations absurdes en sont un bon exemple). De l'autre côté, les approches qualitatives d'observation s'appuient sur des échantillons souvent réduits (de l'ordre de la dizaine voire centaine d'entretiens) pour proposer un schéma explicatif.

(Au moins) une difficulté technique. L'usage des statistiques en sciences sociales n'est pas récent et a même été aux origines de la sociologie avec, par exemple, les travaux d'Emile Durkheim sur les statistiques liées au suicide (1897). Son usage permet fondamentalement de trancher entre des propositions contradictoires sur la compréhension du monde social. Toutefois, il est admis que le travail de modélisation quantitative ne permet pas de calculer de « modèle vrai ». En effet, construire un modèle explicatif (dans le but d'argumenter en faveur d'une explication) nécessite de procéder à une sélection toujours incomplète de facteurs explicatifs (eux-mêmes modélisés à travers des variables comportant une erreur non-nulle, liée à la modélisation et la collecte des données). Aussi, l'articulation des variables au sein d'un système d'équation ou en tout cas d'une structure mathématique symbolisant les phénomènes ciblés est également une réduction du réel. La tâche est ainsi de s'appuyer sur des hypothèses de modélisation (choix des variables, choix de la structure algorithmique, choix du critère de validation) qui permettent de réaliser un test de vérification par la comparaison de ses « sorties calculables » (Varenne, 2008). On différenciera d'ailleurs la construction de modèles prédictifs et des modèles explicatifs. Ces derniers ont une performance prospective plus faible que les premiers car ils doivent remplir une contrainte supplémentaire : devoir à la fois garantir une performance prédictive et rendre compte dans leur structure des phénomènes ciblés (Denis et Varenne, 2019).

Ces quelques remarques permettent de rendre compte sommairement des difficultés ontologiques et épistémologiques qui animent les débats scientifiques sur la modélisation quantitative en sciences sociales. Nous positionnerons notre recherche vis-à-vis de ce débat dans le paragraphe suivant en décrivant le cadre théorique dans lequel ce travail de modélisation s'inscrit.

Cadre théorique

> *« La démarche scientifique est une « arborescence méthodologique qui prend racine dans la posture épistémologique du chercheur et se concrétise par le choix d'instruments de saisie et d'analyse des données, en passant par celui de méthodes, c'est-à-dire de stratégies de recherche visant à mettre au jour des données crédibles en regard de l'objet de recherche » Gohier (2004, p. 3)*

A l'aide d'un travail de synthèse de la littérature, de reformulation et de traduction des travaux portant sur les modèles culturels de Bonnin (2016) et sur le cadre de modélisation de Kowsari (2011), nous proposons de développer un travail de modélisation quantitatif permettant d'étudier la validité des hypothèses sur les systèmes énergétiques domestiques et leur lien avec des situations d'habitation. Le travail de recherche est donc empirique : il s'agit de confronter un modèle théorique à des données d'enquête. Dans le sens, où ce travail cherche à argumenter en faveur d'un déterminisme indiquant que les contextes résidentiels influencent les pratiques sociales et la CED, ce travail s'inscrit dans un paradigme post-positiviste.

La méthodologie de recherche est la modélisation quantitative. Les méthodes quantitatives sont décrites et justifiées dans les différents chapitres de la thèse, de même que les données d'enquêtes mobilisées. Le critère de preuve utilisé dans ce travail de thèse double : il est d'abord quantitatif puisqu'on observe le degré d'homogénéité des classes que l'on construira ainsi que le degré de précision des modèles de régression par rapport aux données collectées. Il est aussi qualitatif car les classes doivent être interprétables. Ce critère de vérification de la théorie n'est pas exclusif : nous considérons que l'expérience permet de discriminer des apports théoriques, mais le degré de précision est un indicateur de la probabilité de proposer un modèle erroné.

3.3 Présentation générale des données

3.3.1 Présentation des enquêtes

Les données seront présentées en détail dans les chapitres où elles sont mobilisées.

Enquête ENERGIHAB

Les données de l'enquête ENERGIHAB sont issues d'une enquête téléphonique menée en 2010 auprès de 1950 ménages de la région Île-de-France. Le questionnaire comportait 362 questions portant sur le comportement, la consommation d'énergie et les caractéristiques des ménages et des logements. Les ménages interrogés ont été sélectionnés pour que des profils hétérogènes de ménages et de logements soient représentés. Les variables ont été construites à partir des réponses aux questionnaires.

Enquête PHEBUS

L'enquête Performance de l'Habitat, Équipements, Besoins et Usages de l'énergie (PHEBUS) a été commandée par le Ministère de la Transition écologique et de la Cohésion des territoires et réalisée en 2013. Cette enquête ponctuelle a été commandée initialement pour disposer d' « une photographie des performances énergétiques du parc des résidences principales » auprès de 8000 ménages représentatifs des régions, des zones climatiques, des types d'habitat et des années de construction.

En reprenant les éléments de présentation de l'enquête sur le site du Ministère, il est indiqué que les données collectées permettent de décrire :

- Les caractéristiques générales du logement et des occupants (taille, date d'achèvement, statut d'occupation…).
- Les caractéristiques socio-démographiques du ménage (composition, âge, sexe, nationalité, diplôme, situation, profession, nature de l'emploi, date d'installation dans le logement, charges, revenus…).
- Les travaux d'amélioration de l'habitat effectués depuis 2008 et pouvant avoir un impact sur l'efficacité énergétique (travaux d'isolation, changement de la chaudière, des fenêtres, installation de panneaux solaires…).
- Les modes de chauffage, les équipements ménagers et automobiles (description des modes de chauffage, des équipements ménagers les plus " énergivores " et des véhicules motorisés).
- Les usages et comportements énergétiques (période de chauffe, réglage de température nuit/jour, pratique d'aération, déplacements…).
- Les consommations d'énergie.

Les données ont été collectée lors d'un entretien direct réalisé par un enquêteur mandaté par l'institut IPSOS. Pour 2000 ménages, un Diagnostic de Performance Energétique (DPE) a été réalisé par un bureau d'étude.

3.3.2 Remarques sur les données

La qualité de ces données, leur suffisance pour réaliser le travail de modélisation proposé sont des éléments critiques pour la qualité du travail de thèse. Nous avons sélectionné les enquêtes ENERGIHAB et PHEBUS pour ce travail, car ces deux enquêtes contiennent des données caractérisant les logements (et le quartier), les ménages (et leur parcours résidentiel), leurs comportements domestiques, et leurs consommations d'énergie. L'enquête ENERGIHAB s'est avérée fournir des données de consommation avec une imprécision importante et l'enquête PHEBUS comporte des données moins riches concernant les comportements domestiques des ménages. Aussi, en ayant bien conscience du caractère réducteur de la mise en variable et de leur imprécision, nous considérons que cette méthode de modélisation permet une approche « tangentielle » du réel cherchant à décrire à la fois des objets-concepts (« situations d'habitation », « systèmes énergétiques domestiques ») et leurs interrelations. Le travail de modélisation réalisé n'est pas perçu comme une observation du réel, mais comme un outil technique intégré dans un travail itératif de construction et d'analyse critique des éléments argumentatifs permettant d'étayer ou d'écarter des hypothèses.

Tableau 5 : Richesse d'information et qualité des données pour chacune des catégories et pour chacune des enquêtes. Source : Auteur.

	ENERGIHAB (2010)	PHEBUS (2013)
Caractères du logement	Riche - bonne qualité	Riche - bonne qualité
Caractères du ménage	Riche - bonne qualité	Moyen - bonne qualité
Pratiques domestiques	Riche - bonne qualité	Moyen - bonne qualité
Consommations d'énergie	Riche et qualité moyenne	Riche - bonne qualité

3.4 Méthodologie

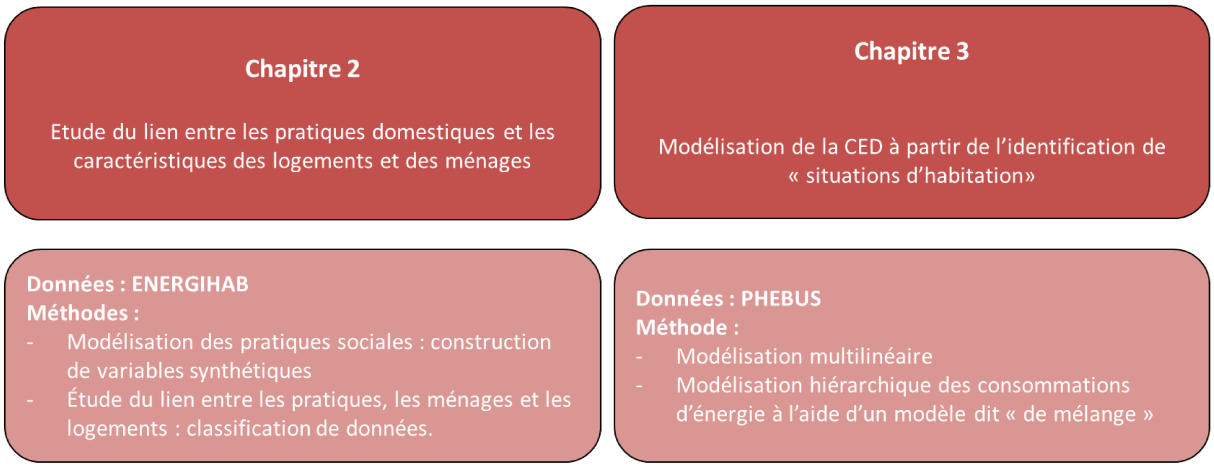
Le modèle pose comme hypothèse centrale l'interaction entre ménage, logement et environnement pour expliquer la construction de la CED. Tout en reconnaissant les effets des facteurs extensifs (surface habité, température de chauffe, résistance thermique des parois, durée d'usage des équipements, nombre d'équipements).

Sur le plan technique, cette approche pose au moins quatre problèmes : (1) le traitement de données qualitatives et quantitatives, (2) la construction d'un modèle (principalement) non linéaire (3) la non-disponibilité de l'ensemble des variables citées dans le modèle, et (4) la forte dépendance du modèle à l'échelle spatiale de l'enquête. Concernant les points (1) et (2) nous proposons dans chacun des chapitres des outils issus de la littérature. Concernant les points (3) et (4), une discussion concernant la stabilité des résultats est menée dans chacune des parties.

La présentation du travail de résultat est structurée de la manière suivante :

- Dans le chapitre 2 nous utilisons les données de l'enquête ENERGIHAB pour étudier le lien entre les caractéristiques des ménages, des logements (les contextes résidentiels) avec les pratiques domestiques et les consommations d'énergie.

- Dans le chapitre 3 nous utilisons les données de l'enquête PHEBUS pour construire un modèle de régression de la consommation qui valorise les liens établis entre les contextes résidentiels et la consommation d'énergie. Nous discutons ensuite de l'influence de chacun des paramètres du modèle et le mettons en regard des résultats clés présentés dans la littérature.



Remarques sur les algorithmes et la démarche de modélisation

La construction d'un « modèle » renvoie dans les faits à faire des hypothèses (nombreuses) pour construire des variables à partir des réponses aux questionnaires, sélectionner un sous-ensemble de la base de données, éventuellement corriger des données, puis les associer dans un algorithme, choisi, puis paramétré, éventuellement modifié par le modélisateur. La description des étapes de modélisation met ainsi en évidence le rôle central du ou des modélisateur(s), ici les membres de l'équipe de recherche. Parmi les hypothèses importantes dans la construction de la stratégie de classification, on citera : la sélection des variables, le pré-traitement des variables (« feature engineering »), la structure de l'algorithme (en particulier sa linéarité, la complexité, le caractère supervisé ou non-supervisé), le critère quantitatif utilisé par l'algorithme pour construire les classes.

Nous soulignons dans ce paragraphe le fait que la démarche de modélisation quantitative suivie dans cette recherche a été itérative (Figure 17) où la formulation d'hypothèses de modélisation a été constamment mise en regard des résultats, permettant ainsi d'améliorer leur robustesse. La thèse contient une synthèse des hypothèses retenues et s'attache, lorsque que cela est nécessaire, à présenter des résultats intermédiaires des travaux de recherche.

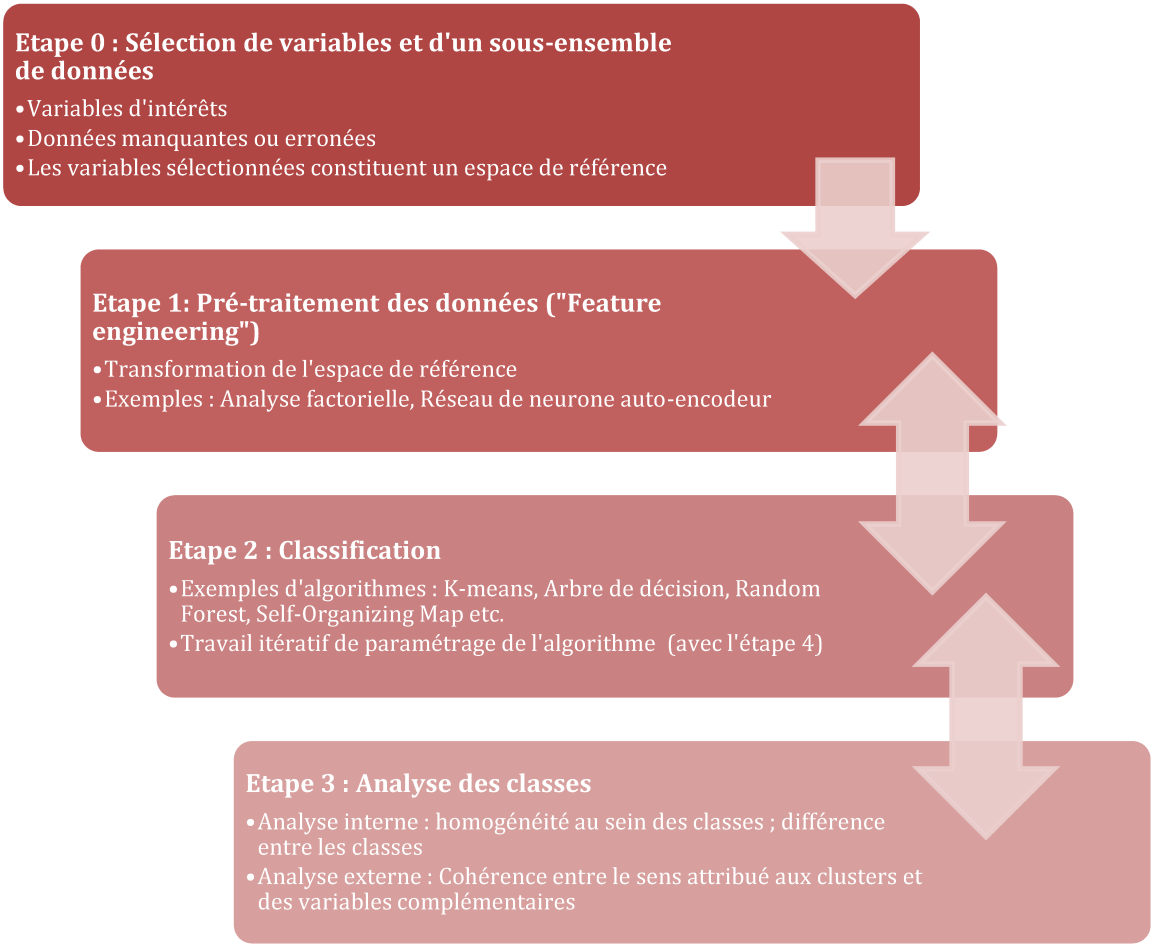


Figure 17 : Liste des étapes de travail dans la construction et l'analyse d'une classification. Les flèches doubles soulignent le caractère itératif du travail de construction d'une classification. Source : Auteur.

Limites techniques

La « validation » empirique du modèle n'est pas réalisable complètement pour plusieurs raisons.

Tout d'abord l'absence de données longitudinales ne permet pas d'étudier la dépendance temporelle des CED avec les ménages et les logements. Nous nous en remettons sur ce point à la littérature existante (Madsen et Gram-Hanssen, 2017 ; Shove, 2017).

Ensuite, ainsi que l'illustrent les travaux de Marguerite Bonnin (2016), la compréhension fine des consommations d'énergie des ménages, et donc la validation empirique du cadre de modélisation proposé, nécessitent d'avoir des données très précises sur les logements et leur environnement (localisation, quartier, climat, ombrages, …) et sur les ménages (parcours résidentiels, états de santé, croyances des individus, rapport au logement, etc.). Même si les ensembles de données se sont enrichis, peu contiennent assez de variables permettant de rendre compte de manière complète des dynamiques qui traversent les situations d'habitations. Dans ce travail de modélisation nous expliciterons chacune des variables explicatives liées aux situations d'habitation pour identifier de quelles interactions nous rendons compte (et par exclusion celles que nous omettrons !).

Sur la Figure 18 on résume l'approche proposée. On considère dans ce travail deux objets : les styles de vies résidentiels, caractérisés par des ensembles équipements et de comportements ; et des situations d'habitation, caractérisés par des ménages et des logements. Leur association permet de déterminer la consommation d'énergie (observable). La caractérisation des deux premiers objets est explorée dans le chapitre 2 et permet ensuite de proposer un modèle dans le chapitre 3.

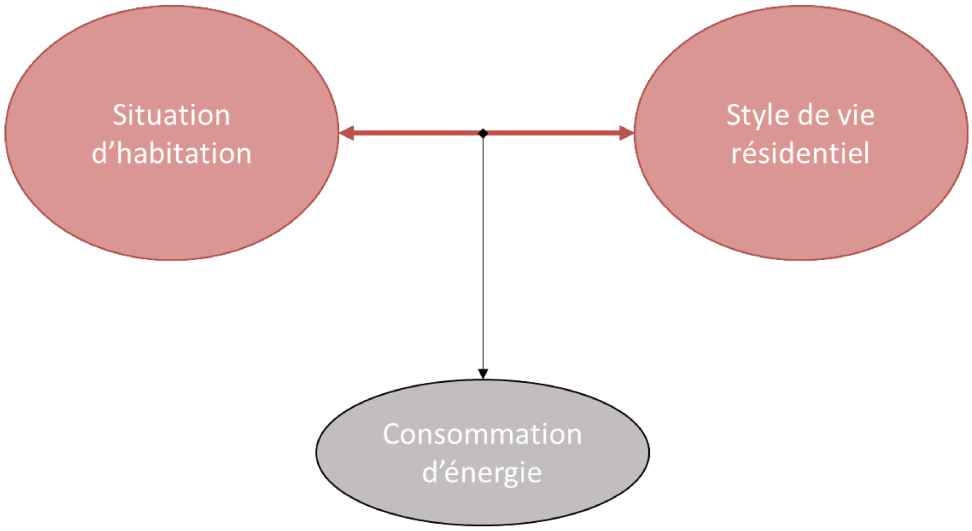


Figure 18 : Schéma du cadre de modélisation simplifié. Source : Auteur.

Chapitre 2. Etude du lien entre pratiques énergétiques domestiques et situations d'habitation

Dans le chapitre 1, nous avons illustré l'intérêt de développer un modèle quantitatif de la CED intégrant les pratiques sociales en lien avec situations d'habitation. Une étape préliminaire consiste à étudier la manière dont les pratiques domestiques, observées à travers des variables construites sur des données d'enquête peuvent être modélisées pour être intégrées dans des modèles de CED. Les pratiques sont effet essentiellement accessibles à travers le croisement des équipements, des attitudes et des comportements des ménages (Shove, 2003) nécessitant la réalisation d'entretiens qualitatifs. Ces données n'étant pas disponibles sur des échelles spatiales étendues, il est nécessaire de procéder à une étape de modélisation, ce qui permet ensuite d'étudier leur liens avec les caractéristiques des logements et des ménages (partie 1). Ensuite, les interrelations entre les pratiques sont étudiées en effectuant un travail de classification des pratiques (partie 2). La littérature ne proposant pas de méthode de référence, une discussion sur la méthode mathématique et la sensibilité des résultats est menée dans la partie 3.

1. Etude mono variée des comportements domestiques

1.1 Présentation de la base de données issue de l'enquête ENERGIHAB

1.1.1 Sélection des variables de comportement

Pour réaliser notre étude il nous a fallu construire des variables à partir des réponses fournies par les enquêtés. Le terme de comportement doit être compris ici au sens large c'est-à-dire qu'il décrit à la fois des comportements « moyens » rapportés par les enquêtés (la température de chauffe moyenne, le nombre de jours d'occupation du logement, etc.) et aussi des équipements possédés. Nous considérons en effet le comportement domestique sur une échelle de temps étendue car nous souhaitons étudier le « système énergétique domestique ». La sélection des variables comportementales est basée sur deux critères.

D'abord les variables de comportement doivent être impliquées dans des processus de consommation d'énergie. Cette implication peut être directe par exemple l'usage d'un équipement, ou impliquer une dimension symbolique comme l'utilisation de lampes LED qui témoigne à la fois d'un usage plus efficace de l'énergie et d'un soin apporté à la régulation de la consommation d'énergie. L'identification de cet ensemble de variables a été fait par « domaines de consommation ». Marguerite Bonnin rappelle dans son travail de thèse que l'usage classique des catégories des « postes de consommation » (chauffage, eau chaude sanitaire, cuisson, appareils électriques) font davantage référence à la comptabilité énergétique dans l'espace domestique qu'aux pratiques énergétiques et aux services requis

par les habitants. Comme Bovay (Bovay, 1987), elle propose de catégoriser les comportements par « domaines de consommation » : l'alimentation, l'hygiène, l'éclairage, la régulation de la température thermique et de la qualité de l'air, les loisirs. Pour chacun de ces postes nous inventorions des comportements traduisant l'équipement, l'intensité d'usage et les comportements de régulation, comme proposé par (Bourgeois, Pellegrino et Lévy, 2017). Nous y adjoignons une catégorie de variables permettant de modéliser le type d'occupation du logement.

Tableau 6 : Liste des variables de comportement construites à partir de l'enquête ENERGIHAB (ANR). Les variables peuvent être catégorielles (C) ou numériques (N). Source : Auteur.

Symbole de la variable de comportement	**Description**	**Type**	**Commentaire**
F_EQ1	Nombre d'équipements alimentaires	N	De 0 à 10 (moyenne = 5 équipements)
F_EQ2	Possession d'un congélateur individuel indépendant	C	2 modalités
F_US1	Nombre de repas pris par semaine à la maison	N	Entre 0 et 14 (moyenne = 9.9)
F_US2	Nombre de jours d'utilisation du four par semaine	N	Entre 0 et 7 (moyenne = 2.6 jours/semaine)
F_US3	Nombre de jours d'utilisation de la plaque de cuisson par semaine	N	Entre 0 et 7 (moyenne = 6 jours/semaine)
F_REG1	Nombre d'appareils consommateurs d'énergie achetés en tenant compte de la classe énergétique	N	Entre 0 et 7 (moyenne = 1 équipement)
HY_EQ1	Nombre d'appareils d'hygiène	N	Entre 0 et 5 (moyenne = 3.7 équipements)
HY_US1	Nombre de douches quotidiennes par jour et par habitant	C	3 modalités
HY_US2	Les membres du ménage prennent des bains	C	2 modalités
HY_US3	Nombre de jours d'utilisation du lave-vaisselle par semaine	N	Entre 0 et 7 (moyenne = 1.9 jours/semaine)
HY_US4	Nombre de jours de lessive à domicile	N	Entre 0 et 7 (moyenne = 3 jours/semaine)
HY_US5	Nombre de jours d'utilisation du sèche-linge par semaine	N	Entre 0 et 7 (moyenne = 0,9 jours/semaine)
HY_REG1	Le ménage fait attention à sa consommation d'eau	C	2 modalités
HY_REG2	Pratique du tri sélectif	C	2 modalités
HY_REG3	Utilisation de produits ménagers « verts »	C	2 modalités
LI_EQ1	Nombre de lampes par mètre carré de surface habitable	N	Entre 8e-3 et 7e-2 (moyenne = 1,2e-1 lumières/m²)
LI_REG1	Proportion de lampes à diodes électroluminescentes (LED)	C	3 modalités
LI_REG2	Présence de lampes halogènes	C	2 modalités
LI_REG3	Niveau de régulation de l'éclairage pour les pièces inoccupées	C	2 modalités
TC_EQ1	Possession d'un système de chauffage d'appoint	C	2 modalités
TC_US1	Température moyenne de chauffage déclarée en hiver	N	Entre 14°C et 38°C (moyenne : 20,3 °C)
TC_US2	Fréquence de l'aération du domicile	N	Entre 0 et 7 (moyenne = 5,9 jours/semaine)
TC_US3	Les bouches d'aération sont parfois obstruées	C	2 modalités
TC_US4	Ouverture des fenêtres pour rafraîchir les pièces	C	2 modalités
TC_REG1	Type de vêtements pendant la période de chauffage	C	3 modalités
TC_REG2	Diminution de la température pendant la nuit ou en cas d'absence	C	2 modalités
TC_REG3	Nombre d'espaces non chauffés en hiver	N	Entre 0 et 5 espaces (moyenne : 0,35)
TC_REG4	Chauffage éteint lors de l'ouverture des fenêtres	C	2 modalités
WL_EQ1	Nombre de téléviseurs, d'ordinateurs personnels, de consoles de jeux et d'appareils informatiques	N	Entre 0 et 12 (moyenne : 3.5)
WL_US1	Nombre moyen d'heures d'utilisation quotidienne du téléviseur principal	N	Entre 0 et 8h (moyenne : 2.5h)
WL_US2	Nombre moyen d'heures d'utilisation quotidienne des principaux ordinateurs personnels	N	Entre 0 et 8h (moyenne : 1.8h)
WL_REG1	Niveau de régulation énergétique des écrans inutilisés	C	3 modalités
OCC1	Niveau de présence pendant les week-ends	C	3 modalités
OCC2	Fréquence des départs en vacances	C	3 modalités
OCC3	Nombre de jours de travail hors du domicile de la PR	C	3 modalités

Ensuite, ces variables doivent pouvoir être extraites de la base de données. Après la sélection des variables, seuls les ménages pour lesquels nous disposons de données de comportement complètes ont été conservés, ce qui donne un tableau de 1363 ménages dont les comportements sont décrits par 35 variables, dont dix-neuf qualitatives et seize quantitatives. Une liste des variables extraites de l'enquête ENERGIHAB est fournie dans le Tableau 5.

1.1.2 Sélection des variables décrivant le contexte résidentiel

En accord avec les hypothèses de modélisation définies dans le chapitre précédant nous utilisons les variables suivantes, que nous extrayons de la table de réponse de l'enquête ENERGIHAB.

Variables caractérisant le logement : la taille du logement (surface), le type de logement, la surface par personne, le statut d'occupation du logement (secteur privé ou public, propriétaire ou locataire), la zone urbaine, la date de construction, le type de chauffage et l'énergie principale.
Variables caractérisant le ménage : âge de la personne de référence (PR), revenus du ménage, statut d'activité de la PR, catégorie socio-professionnelle (CSP) de la PR, le nombre de personnes, le nombre d'enfants, la composition du ménage.
Le tableau suivant (Tableau 7) propose un résumé statistique des variables extraites de la base de données. La base n'est pas représentative des ménages et des logements en Île de France en 2013, en comparaison avec les chiffres fournis par l'INSEE[21]. La proportion de propriétaires est par exemple surestimée (58% au lieu de 48%), tandis que la part de locataires dans le parc public est sous-estimée (13% au lieu de 25%). Une autre différence majeure est que l'INSEE estimait à 72% la part des logements individuels en 2014 tandis que l'enquête utilisée estime à 43% leur part. Cette remarque vise à rappeler que l'enquête avait alors pour but non pas l'estimation de facteurs de régression nécessitant une base représentative, mais plutôt d'étudier la diversité des contextes résidentiels et des pratiques associées.

[21] Pour plus d'information consulter l'article : « Les conditions de logement en Ile-de-France en 2013 », INSEE, accessible à l'URL : https://www.insee.fr/fr/statistiques/1285809

Tableau 7 : Résumé statistique des variables présentant les contextes résidentiels. Les pourcentages sont arrondis à l'unité. Source : Auteur après traitement de la base ENERGIHAB.

Nom de la variable	**Description de la variable et modalité**	**Variables catégorielles** : Proportion en % **Variables numériques** : Moyenne (écart-type) **Taille de l'échantillon** : N = 1363
HH_AGE_C Age de la PR	< 30 ans	36 (3%)
	30-39 ans	180 (13%)
	40-49 ans	298 (22%)
	50-59 ans	306 (22%)
	60-69 ans	234 (17%)
	>70 ans	298 (22%)
	Non connu	11 (1%)
HH_INCOME_C Niveau de revenus du ménage (par quintile)	Q1	271 (20%)
	Q2	131 (10%)
	Q3	299 (22%)
	Q4	181 (13%)
	Q5	123 (9%)
	Non connu	358 (26%)
HH_STATUS_C Statut d'activité de la PR	Active	747 (55%)
	Retraitée	478 (35%)
	Sans emploi	138 (10%)
HH_SPG_C Catégorie socio-professionnelle de la PR	Employés & ouvriers	452 (33%)
	Cadres	289 (21%)
	Agriculteurs et artisans	65 (5%)
	Professions intermédiaires	436 (32%)
	Sans profession	30 (2%)
	Non connu	91 (7%)
HH_NBHAB_N	Nombre de personnes composant le ménage	2.4 (1.4)
HH_COMPO_C Composition du ménage	Couple avec enfant(s)	453 (33%)
	Couple sans enfant	351 (26%)
	Famille monoparentale	98 (7%)
	Personne seule	435 (32%)
	Plusieurs personnes (pas de famille)	26 (2%)
HS_SURF_N	Surface du logement (m²)	91 (59.3)
HS_SURFP_N	Surface du logement par personne (m²/p)	45.7 (29.2)
HS_TYPE_C Type de logement	Logement collectif	783 (57%)
	Logement individuel	580 (43%)
HS_STATUS_C Statut d'occupation	Locataire dans le parc privé	345 (25%)
	Locataire dans le parc public	181 (13%)
	Logé gratuitement	41 (3%)
	Propriétaire ou primo-accédant	796 (58%)
HS_ZONE_C Zone urbaine du logement	Zone périurbaine	129 (9%)
	Zone rurale	522 (38%)
	Zone urbaine	712 (52%)
HS_DATECONSTR_C Date de construction du logement	Avant 1949	336 (25%)
	1949-1975	294 (22%)
	1975-1990	201 (15%)
	Après 1990	195 (14%)
	Inconnu	337 (25%)
HS_HEATING_C Chauffage principal	Chauffage électrique	461 (34%)
	Chauffage au fioul	87 (6%)
	Chauffage au gaz	623 (46%)
	Autre	192 (14%)
HS_CENTRALHEAT_C Chauffage central	Présence d'un chauffage central	641 (47%)
	Aucun	443 (33%)
	Autre	279 (20%)

1.2 Construction de variables synthétiques de comportements et croisement avec les contextes résidentiels

1.2.1 Intérêt de la modélisation des comportements

Etat de l'art et verrous identifiés

L'analyse des liens entre les comportements domestiques recensés dans l'enquête ENERGIHAB et les contextes résidentiels pourrait être effectuée à l'aide de variables simples tels que la possession d'un équipement, le réglage de la température, etc. Toutefois ce type d'exercice présente quelques limites. Tout d'abord, la réponse indiquant la réalisation de gestes comme le réglage de la température, le nombre de repas pris à domicile, la régulation thermique est difficile à interpréter. Les réponses fournies par le répondant peuvent être incomplètes par manque d'information (d'autres personnes peuvent intervenir dans la régulation thermique à son insu), ou ne pas correspondre à la réalité car celle-ci est trop variable pour être moyennée. Enfin le répondant peut ne pas avoir l'information (prenons par exemple le cas d'un répondant n'ayant aucune compétence dans le domaine des pratiques de nettoyage). Enfin, la revue de littérature des enquêtes et des variables extraites dans chacun des cas d'étude de la littérature montre qu'il n'existe pas de méthodologie partagée pour la collecte et l'étude de ces données. La sélection hétérogène des variables et les biais de sélection des échantillons (questionnaires remplis par des étudiants, des universitaires ...) induit des difficultés significatives dans la comparaison des résultats.

Le regroupement des variables de comportement au sein de variables dites « synthétiques » permet par agrégation d'étudier des objets mathématiquement plus robustes à ces sources d'erreur. Dans la littérature, ce travail est effectué soit pour étudier les comportements en eux-mêmes soit pour classifier ensuite les répondants selon les scores qu'ils obtiennent dans cet espace calculé. Il s'agit alors de travaux s'intéressant à la caractérisation de modes de vies résidentiels. Parmi les travaux ayant construit des facteurs synthétiques, on recense notamment les travaux de (Bourgeois, Pellegrino et Lévy 2017) qui ont construit trois indicateurs (niveau d'équipement, intensité d'usage, niveau de régulation) à partir des variables de l'enquête ENERGIHAB et de leur expertise. Ces indicateurs leur permettent ensuite d'étudier le lien entre caractéristiques des ménages et des logements et leurs comportements. Ensuite, plusieurs travaux ont utilisé les variables de comportement au sein d'analyses factorielles pour ensuite construire des typologies de comportement (Ben et Steemers, 2018 ; van Raaij et Verhallen, 1983)[22].

Les hypothèses de travail

Dans ce travail, nous proposons une méthodologie originale pour construire un espace synthétique à partir de données qualitatives et quantitatives. Les méthodes existantes diffèrent selon le critère de regroupement des variables. Les méthodes expertes comme celle de (Bourgeois, 2017) reposent sur une

[22] Voir l'état de l'art réalisé dans le chapitre 1 à la partie 1.2.1 (p. 31).

définition explicite d'indicateurs liée à une connaissance. Les méthodes factorielles quant à elles reposent sur un critère géométrique. L'idée est dans cette perspective que les lignes d'une base de données sont « peu distantes ». Les variables permettant de discriminer les lignes sont identifiées et servent à construire des variables synthétique par combinaison linéaire des variables initiales.

Dans notre travail, nous proposons de discuter cette approche en construisant des variables synthétiques selon un critère de corrélation (*ClustOfVar*). Dans un premier temps, nous présentons la méthodologie et utilisons les résultats de cette classification de variables à partir de la base de données ENERGIHAB. Nous comparons cette classification avec une analyse factorielle de données mixtes (AFDM) dans la partie 2.3 (p.107)

1.2.2 Classification des variables de comportement par la méthode ClustOfVar

Méthodologie

Classification des variables selon un critère de corrélation

L'approche de classification des variables proposée par (Kuentz-Simonet et *al.*, 2013) offre une alternative à l'analyse factorielle pour construire des variables quantitatives à partir d'un mélange de variables quantitatives et qualitatives. Les groupes de variables sont construits sur un critère d'homogénéité (équation 1) qui est défini comme la somme des rapports de corrélation (pour les variables qualitatives) et des corrélations au carré (pour les variables quantitatives) avec une variable quantitative synthétique. La variable synthétique (VS) est calculée comme étant le premier facteur principal calculé par une analyse factorielle des variables appartenant au cluster. Par souci de clarté, le terme "variable synthétique" est repris de la méthode originale. En plus de fournir des variables synthétiques, cette approche permet d'étudier la proximité des variables en termes de corrélation.

$$H\ (C_k) = \sum_{\boldsymbol{x_j} \in C_k} r^2{}_{\boldsymbol{x_j}, \boldsymbol{y_k}} + \sum_{\boldsymbol{z_l} \in C_k} r^2{}_{\boldsymbol{z_l}, \boldsymbol{y_k}}$$

Equation 1 : Homogénéité H du cluster C_k représenté par sa variable synthétique numérique $\boldsymbol{y_k}$. $\boldsymbol{x_j}$ désigne la j^e variable quantitative et $\boldsymbol{z_l}$ désigne la l^e colonne de la table contenant les variables issue du codage disjonctif complet des variables qualitatives.

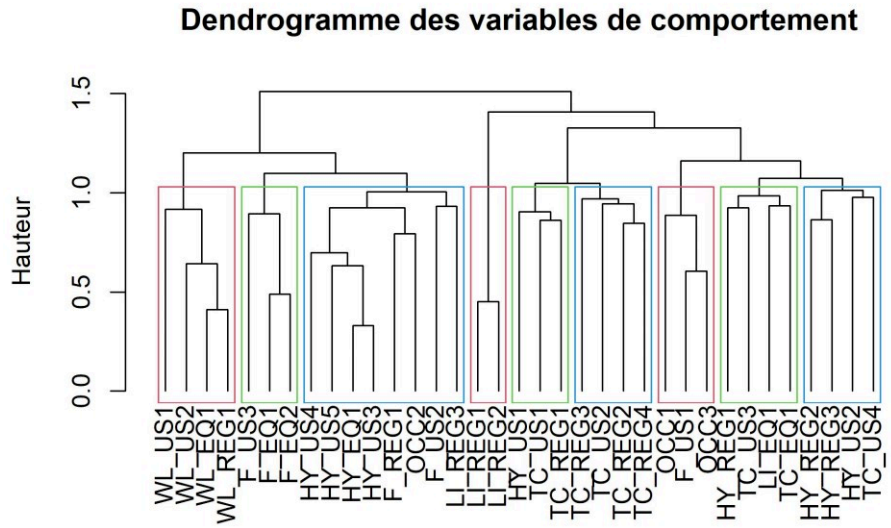
$$d\ (A, B) = H(A) + H(B) - H(A\ U\ B)$$

Equation 2: Dissimilarité d entre les clusters de variables A et B

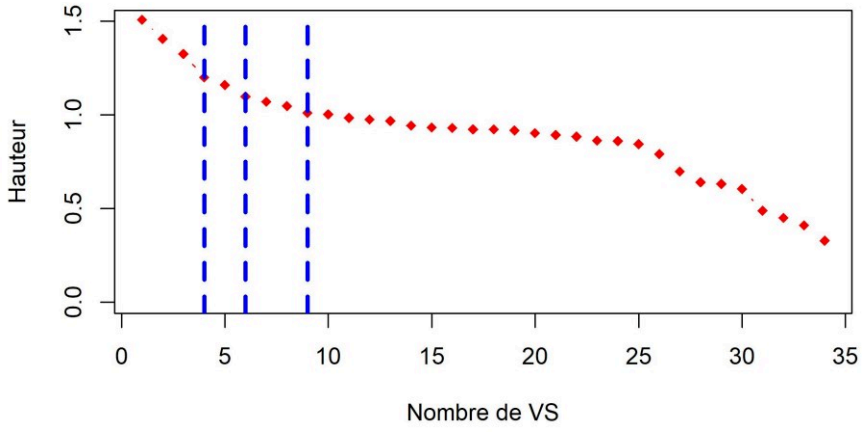
Vocabulaire :

Dans ce travail de thèse, nous utilisons plusieurs termes comme « cluster », « classe », « groupe », « classes », « type ». Ces mots peuvent être considérés comme équivalents : ils proviennent de communautés scientifiques différentes (cluster, classe et classes sont plutôt utilisés dans la communauté en science de la donnée tandis qu'on parlera plutôt de groupe ou de type en dehors).

Le choix du nombre de clusters K_l, c'est-à-dire du nombre de VS est effectué en trois étapes. Tout d'abord, une classification hiérarchique ascendante (CAH) a été utilisée pour construire une structure hiérarchique des variables. L'agrégation a été calculée en minimisant le critère de dissimilarité à chaque étape (équation 2). Ensuite, l'évolution de la hauteur de l'arbre a été visualisée, calculée comme l'accumulation des dissimilarités, en fonction du nombre de clusters de variables. Une première estimation du nombre optimal K_l a été faite en choisissant celui pour lequel la diminution de la hauteur entre les clusters K_l- 1 et K_l était beaucoup plus importante que celle entre les clusters K_l et K_l+ 1. Enfin, nous avons étudié la stabilité de ce choix en comparant les résultats de la même opération de clustering sur 100 sous-ensembles aléatoires de ménages, un processus appelé "bootstrapping". La stabilité est résumée par l'indice Rand ajusté (ARI) qui mesure la récurrence de l'association entre les variables. Dans le cas présent, le critère ARI mesure la persistance des associations entre les variables. Tous les calculs ont été effectués à l'aide du paquet R *ClustOfVar* (Chavent et *al.*, 2012). Un autre exemple de classification de variables utilisant la même méthodologie peut être trouvé dans (Kuentz-Simonet et *al.*, 2013).



(A) Dendrogramme des variables de comportement. La hauteur d'un cluster C associant deux groupes A et B (C= AUB) est définie comme h(C)= d(A,B). Les rectangles en couleur permettent de distinguer les groupes des variables construits pour K_1=9.



(B) Evolution de la hauteur en fonction du nombre de clusters variables.

Figure 19: Dendrogramme (A) et hauteur (B) calculée pour la classification des variables de comportement. Source : Auteur après calculs sur la base ENERGIHAB.

Le regroupement des variables produit une classification hiérarchique des variables de comportement (Figure 19A). Nous avons identifié trois valeurs de K_1 pour lesquelles une forte augmentation de l'homogénéité des clusters a été observée (Figure 19B). Un clustering avec K_1 supérieur à 25 pourrait

être intéressant car il pourrait produire des clusters plus homogènes. Cependant, cela ne répondrait pas au principe de parcimonie. La sélection entre les trois valeurs (K_1 = 4, 6 et 9) a été faite en observant la stabilité et l'interprétabilité des clusters. Puisque nous avons observé que l'indice de Rand ajusté (ARI) augmentait avec K_1 (ARI_{K1}=4 = 38%, ARI_{K1}=6 = 45%, ARI_{K1}=9 = 49%), K_1 =9 a été sélectionné afin de conserver un clustering variable qui dépend moins de l'ensemble de données.

Deuxièmement, nous avons constaté que les clusters avec K_1 =4 ou K_1 =6 produisent des clusters qui mélangent des comportements pour lesquels aucune méthode simple d'interprétation simultanée n'a pu être trouvée dans la littérature. Les clusters de variables avec K_1 =9 sont présentés dans le tableau 2. Un tableau listant les coordonnées et les corrélations utilisées pour l'interprétation et la dénomination des VS est fourni en annexe. L'homogénéité et l'interprétabilité des clusters peuvent être étudiée à l'aide des informations fournies dans le Tableau 8 et du dendrogramme des variables fourni à la Figure 19. On observe que 7 variables initiales sont faiblement corrélées avec la VS (Tableau 7). On remarque que leur contribution, définie comme la corrélation au carré avec la VS est inférieure à 20% (elles sont indiquées par une astérisque). Ces variables n'ont donc pas été considérées pour l'interprétation des VS. Après avoir identifié les variables les plus fortement liées à la VS, la VS peut être interprétée en étudiant les positions des modalités des variables qualitatives sur la VS et en étudiant la corrélation des variables quantitatives avec la VS. Pour faciliter la clarté de la présentation, on présente d'abord un résumé statistique de la classification des variables. Une présentation qualitative des variables synthétiques est proposée au paragraphe 1.2.3.

On constate sur la Figure 19A que les VS agrègent les variables comportementales dans les mêmes domaines de consommation. Par exemple, le cluster situé à gauche du dendrogramme regroupe les variables WL_US1, WL_US2, WL_EQ1, WL_REG1 qui sont des variables liées aux loisirs numériques décrivant respectivement l'intensité d'usage des téléviseurs, des ordinateurs personnels, le nombre d'équipement numériques, et les comportements de régulation des consommation de ces appareils (extinction lorsqu'ils ne sont pas utilisés). Le second groupe regroupe quant à lui des variables décrivant le niveau d'équipement dans le domaine alimentaire (F_EQ1, F_EQ2, F_US3). Aussi, le dendrogramme permet d'observer un ordre de proximité (au sens des corrélations calculées dans la méthode de classification) entre les variables d'enquête : les variables d'équipement sont ainsi ordonnées sur la gauche du dendrogramme, tandis que les variables liées aux usages d'eau chaude sanitaire, au chauffage du logement, à son occupation moyenne sont situées à droite. On retrouve cette organisation dans le Tableau 8 où un nom est donné à chacune des variables synthétiques construites en interprétant les variables regroupées.

Tableau 8: Résumé des liens entre les VS et les variables initiales. Une interprétation des VS est proposée en accord avec les variables initiales identifiées. Source : Auteur après calculs sur la base ENERGIHAB.

Variable synthétique	**VS1**	**VS2**	**VS3**	**VS4**	**VS5**	**VS6**	**VS7**	**VS8**	**VS9**
Interprétation de la variable synthétique	Équipement pour l'alimentation	Présence à la maison	Équipement d'hygiène et utilisation	Niveau de restriction	Besoin en chauffage	Régulation du chauffage	Demande en loisirs	Gestes verts	Type de lampes
Nombre de variables dans le groupe (Degrés de liberté)	3 (3)	3 (6)	8 (10)	4 (6)	3 (6)	4 (5)	4 (5)	4 (8)	2 (5)
% de la variance expliquée par la VS	53,9%	30,2%	29,9%	29,0%	24,7%	31,1%	40,1%	28,8%	51,7%
Variables inclues (corrélation entre la variable et la variable synthétique VS)	F_EQ1 (0,74) F_EQ2 (0,64) F_US3 (0,24)	F_US1 (0,64) OCC3 (0,59) OCC1 (0,28)	HY_EQ1 (0,71) HY_US3 (0,62) HY_US5 (0,42) HY_US4 (0,40) F_REG1 (0,25) OCC2 (0,18)* F_US2 (0,08)* LI_REG3 (0,02)*	HY_REG1 (0,31) TC_US3 (0,3) TC_EQ1 (0,28) LI_EQ1 (0,26)	TC_US1 (0,44) TC_REG1 (0,43) HY_US1 (0,37)	TC_REG4 (0,47) TC_REG2 (0,39) TC_US2 (0,24) TC_REG3 (0,14)*	WL_REG1 (0,72) WL_EQ1 (0,67) WL_US2 (0,48) WL_US1 (0,15)*	HY_REG2 (0,56) HY_REG3 (0,48) HY_US2 (0,06)* TC_US_4 (0,04)*	LI_REG2 (0,77) LI_REG1 (0,77)

Après cette étape, les corrélations entre les variables synthétiques ont été calculées afin de vérifier que chaque VS fournit des informations indépendantes. Les coefficients de corrélation étaient tous inférieurs à 0,25 en valeur absolue, sauf pour les VS 1, 3 et 7. Le coefficient de corrélation entre la VS3 et la VS1 était de 0,42, entre la VS3 et la VS7 de -0,46 et entre la VS1 et la VS7 de -0,34. Cela met en évidence le fait que le nombre d'équipements pour l'alimentation, l'hygiène et le loisir ne sont pas indépendants. Une autre investigation a porté sur la qualité de la représentation des données : bien que l'algorithme vise à construire les variables les plus fortement corrélées, il est intéressant d'observer le pourcentage de variance synthétisée par les variables construites, qui varie entre 24,7% et 53,9%. Cette variation est due au fait que les VS regroupent des variables qui véhiculent une information plus moins similaire. Par exemple, la VS1 regroupe essentiellement des informations sur le nombre d'équipements alimentaire (F_EQ1) et la possession ou non d'un congélateur indépendant (F_EQ2). Il nous semble normal que cette variable quantitative VS1 garde un pourcentage d'information plus élevé (53,9%) qu'une variable quantitative synthétisant des pratiques de régulation du chauffage (VS6 – 31.1%) qui sont plus diverses.

1.2.3 Croisement des variables synthétiques avec les contextes

Nous avons construit des variables synthétiques à l'aide d'un critère de corrélation. Dans cette partie nous étudions la distribution de ces variables et leurs liens avec les caractéristiques des contextes résidentiels. Nous utilisons dans cette partie une analyse linéaire à un facteur pour étudier la liaison entre les variables synthétiques avec les autres variables. Nous modélisons linéairement le lien entre la i^{e} variable synthétique et la variable x. Dans le cas où la variable x est numérique, on pose le modèle de régression linéaire :

$$VS_i = \alpha_0 + \alpha_1 x + \epsilon$$

Dans le cas où x est une variable qualitative, nous nous plaçons dans le cadre de l'analyse de la variance (ANOVA)

$$VS_i = \mu + \tau_k + \epsilon$$

Nous pouvons ainsi extraire les effets α_1 des variables quantitatives et τ_k des variables qualitatives, ainsi que la probabilité de commettre une erreur en jugeant de la non-nullité de ces effets (p-valeur). Cette probabilité est estimée par un test de Fisher pour les variables qualitatives et par un test de Student pour les variables quantitatives. Les résultats de ces modélisations sont synthétisés dans le Tableau 9. Une analyse des résultats est donnée ci-dessous.

Equipements alimentaires et intensité usage (VS1)

Le niveau d'équipement alimentaire semble être lié au revenu du ménage, à la composition du ménage et son nombre d'habitants, à la surface du logement, au statut d'occupation et aux variables VS3, VS7, VS8 et VS9. L'interprétation des effets individuels de ces variables semble montrer que le revenu

semble avoir un effet positif sur le niveau d'équipement alimentaire, de même que le fait d'être propriétaire et d'habiter dans un logement individuel, et le fait que le ménage comprenne des enfants.

Présence au logement (VS2)

La présence au logement semble être liée à l'âge de la PR, au revenu, à la catégorie socio-professionnelle de la PR, au statut d'activité de la PR, au type et à l'âge du logement, ainsi qu'aux variables VS4, VS5, SV6, SV7, VS9. Comme attendu, un âge plus élevé, un statut d'inactif accroissent le niveau de présence au logement. On observe également qu'un revenu plus élevé diminue la présence au logement.

Equipements d'hygiène et intensité d'usage (VS3)

Le niveau d'équipement et l'intensité d'usage des équipements d'hygiène possèdent des liens similaires à la variable VS1 avec les autres variables. Liée à l'âge de la PR, au revenu, à la CSP, au statut d'activité, au statut d'occupation, à la zone urbaine, au type de logement et sa taille, la VS3 est également liée aux variables VS1, VS6, VS7, VS8, VS9.

Niveau de restriction (VS4)

Cette variable de comportement regroupe des comportements de restriction (économie d'eau, de chauffage, usage préférentiel d'appareils auxiliaires de chauffage). Elle est liée au revenu du ménage, au statut d'activité, à la taille et au type de logement. Elle est aussi liée plus marginalement aux variables SV5 et SV8. L'étude des effets nous montre que les comportements de restriction sont plus présents chez les personnes aux revenus plus faibles, inactives et dans des logements collectifs. Ils sont aussi liés à une température de chauffe moyenne déclarée plus faible (VS5) ainsi qu'à des gestes de régulation (VS8) plus importants que la moyenne.

Besoin en chauffage (VS5)

La variable VS5 est construite à partir de la température moyenne de chauffe déclarée et les habitudes de porter des vêtements chauds en période hivernale. Cette variable est liée à très peu de variables des contextes résidentiels. Elle est liée à la CSP de la PR, à l'âge du logement et au type de chauffage. Les pratiques de chauffage semblent être moins intensives dans les logements très anciens, chauffés à l'électricité et avec un chauffage individuel. En revanche la variable semble être liée aux variables synthétiques VS2, VS4, VS6, VS7, VS9. La liaison le plus forte semble être entre la régulation du chauffage (VS6) et VS5 : des pratiques plus élevées de régulation semblent être anti-corrélées avec un température de chauffage importante. La faible liaison de cette variable avec les contextes résidentiels pourrait s'expliquer d'une part par le fait que les températures moyennes de chauffe déclarées sont souvent erronées parce que méconnues ou parce que la réponse fournie par les enquêtés peut être influencée par un biais de désirabilité sociale en rapprochant la réponse de la norme bien connue des 19°C. Une autre explication probable est que le modèle culturel de chauffage soit peu dépendant du

contexte résidentiel, sinon des leviers qui permettent ou non son contrôle (type de chauffage, contrôle de l'aération, qualité de l'isolation, connaissance et compréhension des systèmes de gestion).

Régulation du chauffage (VS6)

A contrario, la variable VS6 qui synthétise des gestes de régulation du chauffage (extinction durant des périodes d'absence ou lors de l'aération, nombre de pièces non chauffées) est liée à plusieurs variables du contexte résidentiel, dont l'âge de la PR, le revenu, le statut d'activité, la CSP, le type de logement, la zone urbaine, l'âge du logement et le caractère collectif du chauffage principal. Ces pratiques de régulation semblent d'autant plus importantes que la PR est jeune. Les ménages aux revenus moyens semblent avoir les gestes de régulation les plus importants, de même que les cadres ou les professions intermédiaires. Le fait d'être locataire dans le parc public diminue en revanche l'importance de ces gestes.

Demande en loisirs (VS7)

La variable VS7 traduit le niveau d'équipement électroniques et leurs usages. Cette variable est liée à toutes les autres variables à l'exception du type de chauffage et les variables VS4 et VS8. On observe que l'âge de la PR est corrélée négativement avec le niveau d'équipement. En termes de revenus, seuls les revenus les plus bas (premier quintile) ont des niveaux d'équipement faibles. VS7 est également corrélée positivement avec le nombre d'habitants, le fait d'être propriétaire, la surface du logement, et l'âge du logement. On remarque enfin que ce niveau d'équipement est corrélé positivement avec les niveaux d'équipement dans les domaines de l'alimentation et de l'hygiène.

Gestes verts (VS8)

Les gestes de régulation comme la pratique du tri, l'usage de produits ménagers « verts » sont corrélés avec le statut de propriétaire, un âge de la PR supérieur à 40 ans, au nombre de personnes composant le ménage, le fait d'occuper une maison individuelle. Enfin ces gestes « verts » sont corrélés positivement avec le niveau d'équipement.

Types de lampes (VS9)

La variable SV9 caractérise le niveau et la qualité de l'éclairage. On observe que les ménages très jeunes ou très âgés sont plus souvent équipés avec des lampes halogènes. Le nombre de personnes et la présence d'enfants sont des facteurs corrélés avec l'usage de lampes LED, contrairement au revenu. Aussi, l'usage d'équipements LED est corrélé avec un équipement globalement plus important dans le logement. Enfin, l'usage de LEDs est lié positivement avec un besoin de chauffage plus important mais également des gestes de régulation plus nombreux.

L'étude des relations entre les variables de comportements et les variables caractérisant le ménage et le logement est intéressante : elle permet d'observer un lien empirique entre les variables et (re)trouver des liens (par exemple le lien entre niveau d'équipement et revenu). Cette analyse trouve cependant ses limites lorsqu'il s'agit de comprendre la logique habitante qui lie ces caractéristiques. On peut alors s'intéresser aux liens entre les pratiques énergétiques domestiques et les contextes résidentiels. Comme nous l'avons observé dans la littérature, si les comportements sont liés aux contextes résidentiels ils semblent aussi liés entre eux, au sein de ce qui est qualifié parfois de « modes de vies résidentiels ». Cette analyse croisée des comportements et des contextes résidentiels peut être réalisée à l'aide d'un travail de classification de données, que nous explorons dans la suite de ce chapitre.

Tableau 9 : Tableau de synthèse présentant les résultats de la régression simple des variables synthétiques avec les caractéristiques des ménages et des logements. On recense ici les effets et la probabilité d'erreur du test de nullité des effets estimés. Ainsi, une p-valeur proche de 0% indique la probabilité de ne pas se tromper en énonçant que la VS dépend tendanciellement de la variable/modalité citée dans la ligne. Le code couleur permet de repérer les p-valeurs proches de zéro. Source : Calculs de l'auteur à partir des données ENERGIHAB.

	VS1		VS2		VS3		VS4		VS5		VS6		VS7		VS8		VS9	
	Équipement pour l'alimentation		Présence à la maison		Équipement d'hygiène et utilisation		Niveau de restriction		Besoin en chauffage		Régulation du chauffage		Demande en loisirs		Gestes verts		Type de lampes	
	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur	Effet	P-valeur
Age de la PR (Référence 30-39 ans)																		
<30 ans	0,19	41%	-0,18	33%	0,53	6%	-0,17	38%	0,35	8%	-0,30	14%	0,13	58%	-0,69	0%	0,60	1%
40-49 ans	-0,34	0%	-0,18	5%	-0,41	1%	-0,19	6%	0,11	29%	-0,23	2%	-0,05	72%	-0,01	92%	0,08	47%
50-59 ans	-0,16	18%	-0,26	1%	-0,08	58%	-0,09	39%	0,16	12%	-0,20	5%	-0,26	4%	0,04	68%	0,26	3%
60-69 ans	-0,12	34%	-1,27	0%	0,24	12%	-0,06	57%	0,04	74%	-0,35	0%	-0,74	0%	0,05	62%	0,42	0%
>70 ans	0,16	19%	-1,81	0%	0,94	0%	0,01	95%	-0,10	34%	-0,72	0%	-1,32	0%	0,03	80%	0,65	0%
Non connu	0,62	11%	-0,94	0%	0,69	16%	-0,18	59%	0,20	56%	-0,32	35%	-0,80	5%	-0,34	30%	0,35	36%
Quintile de revenus du ménage (référence Q1)																		
Q2	-0,35	1%	0,51	0%	-0,77	0%	-0,06	62%	-0,03	83%	0,22	6%	0,96	0%	0,06	61%	-0,13	31%
Q3	-0,79	0%	0,53	0%	-1,18	0%	-0,14	12%	0,02	83%	0,33	0%	1,12	0%	0,01	87%	-0,14	18%
Q4	-0,87	0%	0,59	0%	-1,99	0%	-0,20	5%	-0,12	25%	0,42	0%	1,56	0%	0,13	21%	-0,14	26%
Q5	-0,97	0%	0,78	0%	-2,64	0%	-0,34	0%	0,00	100%	0,23	6%	1,94	0%	0,15	18%	-0,16	25%
Non connu	-0,46	0%	0,17	8%	-1,00	0%	-0,27	0%	-0,09	33%	0,15	10%	0,74	0%	0,05	53%	-0,05	59%
Statut d'activité de la PR (Référence : Actif)																		
Retraité	0,23	0%	-1,88	0%	0,80	0%	0,15	1%	-0,14	4%	-0,43	0%	-1,05	0%	0,05	43%	0,38	0%
Sans emploi	0,03	78%	-1,79	0%	0,21	16%	0,06	54%	-0,03	73%	-0,15	14%	-0,40	0%	-0,17	10%	0,15	19%
Catégorie socio-professionnelle (Référence : Ouvriers et employés)																		
Cadres et dirigeants d'entreprise	-0,09	34%	0,26	0%	-0,75	0%	-0,03	75%	-0,19	2%	0,24	0%	0,53	0%	0,09	28%	0,07	48%
Agriculteurs et artisans	0,10	54%	-0,14	38%	-0,41	5%	-0,24	9%	-0,15	32%	0,05	74%	-0,24	20%	0,25	8%	-0,05	78%
Professions intermédiaires	-0,05	59%	0,34	0%	-0,64	0%	0,00	99%	-0,19	1%	0,31	0%	0,41	0%	0,17	2%	0,01	90%
Pas de profession	0,29	23%	-0,55	1%	0,61	5%	-0,23	26%	-0,33	11%	-0,45	3%	-0,77	0%	0,25	22%	0,17	48%
Non connu	-0,11	45%	-0,79	0%	-0,69	0%	-0,05	71%	-0,18	16%	0,13	29%	0,38	2%	-0,16	19%	-0,12	39%
Nombre de personnes composant le ménage	-0,30	0%	0,08	0%	-0,60	0%	-0,05	2%	0,00	90%	0,03	21%	0,42	0%	0,06	1%	-0,12	0%
Composition du ménage (Référence : Famille avec enfants)																		
Couple sans enfants	0,34	0%	-0,58	0%	0,96	0%	0,02	75%	-0,01	93%	-0,03	71%	-0,81	0%	0,01	86%	0,35	0%
Famille monoparentale	0,46	0%	0,15	27%	1,29	0%	0,03	79%	0,10	40%	-0,24	5%	-0,51	0%	-0,11	35%	0,22	11%
Personne seule	1,11	0%	-0,33	0%	2,20	0%	0,18	2%	-0,05	48%	-0,13	8%	-1,55	0%	-0,23	0%	0,49	0%
Plusieurs personnes (pas de famille)	0,53	3%	-0,22	36%	1,18	0%	0,16	46%	-0,11	64%	0,00	99%	-0,85	0%	-0,06	77%	0,51	4%
Surface du logement	-0,01	0%	0,00	23%	-0,01	0%	0,00	0%	0,00	93%	0,00	4%	0,01	0%	0,00	0%	0,00	11%
Surface par personne	0,00	21%	-0,01	0%	0,00	9%	0,00	0%	0,00	66%	0,00	47%	-0,01	0%	0,00	10%	0,00	1%
Type de logement (Référence : Collectif)																		
Individuel	-0,88	0%	-0,23	0%	-1,03	0%	-0,13	3%	-0,11	6%	0,27	0%	0,41	0%	0,31	0%	-0,09	19%
Statut d'occupation du logement (Référence : Locataire dans le parc privé)																		
Locataire dans le parc public	-0,25	3%	0,01	93%	-0,19	18%	0,16	11%	0,20	5%	-0,39	0%	0,05	70%	0,09	38%	-0,10	37%

Logé gratuitement	-0,11	58%	0,25	21%	-0,61	2%	-0,38	3%	0,22	23%	-0,13	48%	0,42	7%	-0,18	31%	-0,29	16%
Propriétaire et accédants	-0,68	0%	-0,29	0%	-0,96	0%	0,01	92%	-0,10	16%	-0,05	46%	0,34	0%	0,31	0%	0,05	50%
Zone urbaine (Référence : Zone périurbaine)																		
Zone rurale	-0,47	0%	-0,05	65%	-0,88	0%	0,14	20%	-0,20	6%	0,41	0%	0,63	0%	-0,07	49%	-0,09	47%
Zone urbanisée	-0,06	61%	-0,08	49%	-0,49	0%	0,12	24%	-0,12	24%	0,10	36%	0,54	0%	-0,24	2%	0,16	17%
Date de construction du logement (Référence : 1949-1975)																		
1975-1990	-0,18	12%	0,23	4%	-0,37	1%	-0,02	86%	-0,06	57%	0,53	0%	0,52	0%	-0,07	46%	0,01	90%
>1990	0,02	87%	0,36	0%	-0,48	0%	-0,10	33%	-0,06	56%	0,64	0%	0,55	0%	-0,05	59%	-0,10	40%
<1949	0,10	31%	0,06	51%	0,14	28%	0,15	8%	-0,34	0%	0,24	1%	0,04	73%	0,00	98%	0,19	6%
Non connu	0,38	0%	0,33	0%	0,55	0%	-0,09	32%	0,06	47%	0,17	5%	-0,01	92%	-0,23	1%	-0,01	90%
Type de chauffage principal (référence : Chauffage électrique)																		
Fioul	0,14	33%	0,04	80%	-0,17	38%	0,04	77%	0,31	2%	-0,25	6%	-0,16	35%	-0,04	77%	-0,08	60%
Gaz	-0,03	74%	-0,11	14%	-0,15	13%	-0,02	81%	0,09	18%	-0,11	11%	-0,05	54%	0,05	42%	-0,05	50%
Autre	-0,11	32%	-0,13	23%	-0,16	26%	0,03	77%	0,18	6%	0,02	80%	-0,01	93%	0,20	3%	-0,01	95%
Chauffage individuel (Référence : Oui)																		
Non	0,15	6%	0,09	25%	-0,02	85%	0,08	23%	-0,16	2%	0,38	0%	0,04	69%	0,03	63%	-0,13	9%
Autre	0,38	0%	0,05	54%	0,46	0%	0,13	9%	0,39	0%	-0,71	0%	-0,24	2%	-0,09	24%	0,00	98%
Variables synthétiques de comportement																		
VS1			0,00	99%	0,55	0%	0,01	53%	-0,05	5%	-0,03	20%	-0,38	0%	-0,07	0%	0,08	0%
VS2	0,00	99%			-0,13	0%	-0,05	4%	0,09	0%	0,12	0%	0,28	0%	-0,01	79%	-0,08	0%
VS3	0,33	0%	-0,07	0%			0,02	23%	-0,01	43%	-0,08	0%	-0,40	0%	-0,09	0%	0,06	0%
VS4	0,02	53%	-0,07	4%	0,05	23%			-0,07	1%	0,04	14%	-0,03	45%	0,06	3%	0,02	52%
VS5	-0,06	5%	0,11	0%	-0,03	43%	-0,07	1%			-0,15	0%	0,12	0%	-0,04	9%	-0,07	2%
VS6	-0,04	20%	0,15	0%	-0,16	0%	0,04	14%	-0,15	0%			0,10	1%	0,03	20%	-0,06	3%
VS7	-0,30	0%	0,21	0%	-0,53	0%	-0,02	45%	0,07	0%	0,06	1%			0,00	99%	-0,11	0%
VS8	-0,09	0%	-0,01	79%	-0,20	0%	0,06	3%	-0,05	9%	0,04	20%	0,00	99%			-0,12	0%
VS9	0,09	0%	-0,08	0%	0,10	0%	0,01	52%	-0,06	2%	-0,05	3%	-0,14	0%	-0,09	0%		

2. Construction d'une typologie de « styles de vies résidentiels »

Dans cette partie, nous réalisons un travail de classification de données de comportements afin d'étudier les relations entre les pratiques. Ce travail de classification s'ancre dans une littérature avec une double identité : mathématique d'abord parce que ce type de travail s'appuie sur des algorithmes de classification de données et sociologique ensuite parce que les travaux cherchent à analyser et comprendre les logiques habitantes à partir des résultats des calculs de classification.

2.1 Positionnement et méthodologie

Verrous scientifiques

La construction de types de comportements n'est pas récente. On a effectué une revue de littérature des travaux de classification des données de comportement dans le chapitre 1 (voir p. 32). Ces travaux ont cherché à construire des typologies de comportements de ménages avec différentes motivations (segmentation de la population, aide à la décision publique, simulation prospective de la CED). Dans ce travail de thèse, nous cherchons à produire un modèle de consommation qui permette de mieux comprendre et d'expliquer la consommation d'énergie domestique. En réduisant la complexité de lecture du tableau de données, qui comporte souvent plusieurs centaines voire milliers de lignes, le partitionnement de données de comportements sur des échantillons permet un bon équilibre entre la fidélité des données produites et la formulation d'hypothèses sur le comportement.

L'état de l'art permet cependant de mettre en avant plusieurs limites :

Un nombre important de travaux ne s'intéressent qu'à un nombre réduit de comportements domestiques comme la régulation de la température, l'usage de l'eau chaude sanitaire, régulation de l'usage des luminaires etc.

Peu de travaux de classification étudient la sensibilité de leurs travaux à leur méthode de partitionnement. En particulier, l'analyse factorielle couplée à une classification hiérarchique occupe une place prépondérante parmi les méthodes de segmentation, et ce en dépit des limites connues de l'analyse factorielle pour la construction de typologies. En particulier, De Soete et al. note que l'espace des composantes principales orthogonales ne permet pas toujours la représentation de l'information taxonomique (De Soete et Carroll, 1994).

Il ne semble pas exister de méthodologie partagée, ce qui limite les possibilités de comparaison des résultats (sélection et construction de variables, algorithme, analyse des résultats). On notera par exemple, le fait que la sélection, le codage et éventuellement le seuillage des variables sont rarement explicités alors que leur influence sur le résultat est importante. Aussi, le critère de sélection (définition de la distance, de l'algorithme de segmentation) est peu discuté.

Aussi, si beaucoup de travaux quantifient le lien entre les types de comportements construits et les caractéristiques des ménages et des logements, peu de travaux les étudient en croisant ces caractéristiques, par exemple en étudiant les profils types des ménages et des logements. La construction de profils de contextes résidentiels permet selon nous de mieux comprendre les types de comportements construits.

Enfin, peu de travaux de classification sur les styles de vies résidentiels discutent de la difficulté à manipuler des données hétérogènes associant variables quantitatives, qualitatives (nominales et ordinales). Par souci de simplicité, nous appellerons ces données des « données mixtes ».

Mode de vie ou style de vie ?

Une seconde difficulté de vocabulaire réside dans le choix entre les terminologies de « mode de vie », de « style de vie » et de « genre de vie », qui sont régulièrement mobilisées dans les littératures s'intéressant aux consommations d'énergie. L'article « Mode de vie : de quoi parle-t-on ? Peut-on le transformer ? » de Bruno Maresca (Maresca 2017) permet d'apporter quelques éléments importants pour diriger notre travail. Il note d'abord que le « mode de vie » a très souvent le statut de pré-notion et qu'elle est très peu discutée, même parmi les grands auteurs. Le néophyte est ainsi réduit à observer l'usage qui en est fait dans la littérature en sociologie, qui « suggère donc de penser le mode de vie [au singulier] comme une structure, un cadre de référence, de la vie sociale et de regarder les modes de vie [au pluriel] comme des ensembles de pratiques de la vie quotidienne déclinables à différentes échelles (catégories sociales, territoires, individus) ». En revanche, le « style de vie » renvoie à une logique de différenciation individuelle. En reprenant une généalogie des usages des deux terminologies, l'auteur montre que c'est au XX^{e} avec la formalisation des sciences humaines que se sont construites les racines des 3 termes. Le « genre de vie » renvoie ainsi aux travaux de la géographie et d'anthropologie qui ont décrit « l'influence de l'environnement dans la compréhension de ce qui différencie les peuples et leurs modes d'organisation, notamment matériels. ». Ainsi, avec le développement de l'idée durkheimienne de structures sociales structurant les comportements des individus puis l'idée d'une diversité des trajectoires de socialisation, ce sont en fait les échelles macro-sociales et micro-sociales qui ont été en réalité définies respectivement par le « genre de vie » (société rurale, primitive), et le « style de vie ». Le mode de vie est ainsi un concept « resté dans le flou sémantique », situé entre l'échelle de l'environnement et du collectif et l'échelle de l'individu. Dans la suite de l'article, Bruno Maresca propose une « rigidification » de cette définition afin d'en faire un concept opérant dans le cadre de la transition écologique, en redéfinissant le mode de vie comme étant un système à la fois structurant et différenciant. Les modes de vie seraient ainsi articulés avec les styles de vie où seraient produites les innovations des individus. A ce point de la réflexion, il apparaît que dès lors que notre intérêt se limite à l'étude d'un ensemble de comportements résidentiels, il apparaît plus rigoureux et pertinent d'utiliser la terminologie de « styles de vies résidentiels ». En effet, en accord avec le modèle de Kowsari adopté pour ce travail, ce terme permet

de rendre compte à la fois des déterminismes sociaux et techniques qui contraignent les individus et des appropriations et des innovations des individus.

Dans cette partie nous souhaitons explorer plusieurs méthodes de classification de données pour construire une typologie des comportements résidentiels. Ce travail poursuit ainsi deux objectifs :

- Construire une typologie qui permette de comprendre les liens entre les styles de vies résidentiels et les contextes résidentiels où ils naissent.
- Étudier la dépendance des résultats à l'algorithme et contribuer à construire une méthodologie de classification de ces données mixtes.

Méthodologie

Dans ce travail nous comparons 4 stratégies de classification présentant chacune des avantages et des inconvénients. Ces stratégies sont synthétisées dans le Tableau 10.

Tableau 10 : Synthèse de la problématique de travail et des 4 stratégies de modélisation proposées. Source : Auteur.

Problématique

- Classification des variables de comportement dans les logements
- Identification de styles de vie résidentiels
- Stabilité des résultats, dépendance des résutlats à la méthode de classification

Données

- 35 variables de comportements concernant 1363 ménages de la région Île de France (Enquête ENERGIHAB 2010)
- 14 variables complémentaires décrivant le contexte résidentiel

Stratégie S1: Gower (GOW)+ CAH

- Calcul d'une matrice de distances entre individus par la distance de Gower (GOW)
- Classification par CAH

Stratégie S2 : Analyse Factorielle de Données Mixtes (AFDM) + CAH

- Classification des variables par analyse factorielle de données mixtes (AFDM)
- Classification des scores par classification ascendante hiérarchique (CAH)

Stratégie S3: ClustOfVar (COV) + CAH

- Classification des variables par la méthode ClustOfVar (COV) - (Chavent, 2010)
- Classification des scores par CAH

Stratégie S4: Co-Clustering de données mixtes (COC)

- Co-clustering des données mixtes par la méthode MixedClust (Selosse, 2020)

Le choix de ces stratégies de classification a été réalisée selon la logique suivante :

- La stratégie S1 se base sur l'idée que pour construire une typologie, il est possible de commencer par définir une distance entre les lignes de la base de données. En l'occurrence, la base contenant des données mixtes, il est possible d'utiliser la distance de Gower qui est la distance la plus utilisée dans littérature dans ce cas. Ensuite, le partitionnement de la base de données est fait par

classification ascendante hiérarchique (CAH) car c'est une méthode de référence, facile à implémenter, très visuelle et explicable.

- La stratégie S2 propose de discuter le choix de la distance fait dans la stratégie S1. Par hypothèse, la distance dépend du nombre et des variables sélectionnées, ce qui fragilise la typologie construite. Une manière de rendre la classification plus robuste est alors d'agréger les variables dans des variables synthétiques. L'analyse factorielle de données mixtes (AFDM) permet d'en construire (on les appelle composantes principales) par combinaison linéaire des variables initiales et selon un critère de variance. Une fois ces variables construites, il devient possible de calculer une distance entre les lignes de la base de données en calculant la distance euclidienne dans l'espace synthétique, puis de réaliser une segmentation par CAH.
- La stratégie S3 reprend la critique de la distance de Gower et l'idée de la construction de variables synthétiques de la stratégie S2. Elle diffère de cette dernière en soulignant le fait que l'espace synthétique créé par la méthode AFDM n'est pas un espace contenant l'information taxonomique. La stratégie S3 propose ainsi de s'appuyer sur une autre méthode pour construire l'espace synthétique. La méthode proposée par (Chavent, 2012) procède en classifiant les variables de comportements selon un critère de corrélation. Une fois cet espace synthétique créé, on procède à nouveau à une classification ascendante hiérarchique.
- La dernière méthode S4 propose quant à elle de discuter le choix de définir une distance entre les lignes de la base de données pour construire une typologie. En effet, définir une distance que ce soit à partir des variables initiales ou dans un espace synthétique fait peser des choix méthodologiques lourds sur les résultats. Une alternative peut être alors de considérer la proximité entre des distributions de probabilité. Dans ce cas on fait l'hypothèse que les variables sélectionnées dans la base de données sont assimilables à des variables aléatoires dont la distribution de probabilité est par exemple une distribution gaussienne pour les variables quantitatives ou multinomiale pour les variables qualitatives. Avec cette hypothèse, on peut calculer les paramètres lois qui modélisent le phénomène statistique sous-jacent. Avec cette modélisation, on peut alors comparer les distributions entre les variables et associer au sein d'un même groupe de variable les variables qui ont des distributions proches, puis associer les lignes de la base de données qui ont les mêmes niveaux. Le critère de construction des clusters de lignes et de colonnes devient alors un critère probabiliste : c'est la probabilité d'observer les données de la base, au regard du modèle calculé qui permet de qualifier la qualité de la modélisation. Dans notre calcul, nous utilisons la méthode MixedClust (Selosse, 2020).

Les forces et faiblesse méthodologiques et techniques des méthodes sont synthétisées à la Figure 20.

Figure 20 : Liste des méthodologies de classification de données utilisées pour construire une typologie de styles de vies résidentiels. (Source : Auteur).

Stratégie	Forces	Faiblesses
S1 **GOW + CAH**	• Méthode très connue • Simplicité de mise en œuvre • Pas de prétraitement des données, hormis leur normalisation	• Forte dépendance des résultats à la sélection des variables • Choix de K (nombre de classes) • Pas de variable synthétique de comportement
S2 **AFDM + CAH**	• Simplicité de mise en œuvre • Méthode très connue • Utile pour la représentation des données	• Interprétation des axes factoriels pas toujours aisé • Le critère de construction de l'espace synthétique ne permet pas de préserver l'information taxonomique • Choix de K_1 et K_2 (classification des variables et des lignes)
S3 **COV + CAH**	• Simplicité de mise en œuvre • Relativement bonne interprétabilité des variables synthétiques générées	• Chaque groupe de variables est résumé par une seule variable synthétique, indépendamment de la complexité inhérente au groupe de variables. • Les VS sont générées comme combinaisons linéaires des variables initiales. • Choix de K_1 et K_2 (classification des variables et des lignes)
S4 **COCL**	• Critère probabiliste pour la classification. S'affranchit de la notion de « distance » entre répondants.	• Impossibilité de regrouper des variables quantitatives et qualitatives au sein d'un même cluster de variable • Forte dépendance du résultat à l'initialisation : convergence vers des optimums locaux. • Nombreux paramètres et donc gourmand en ressources computationnelles et en temps.

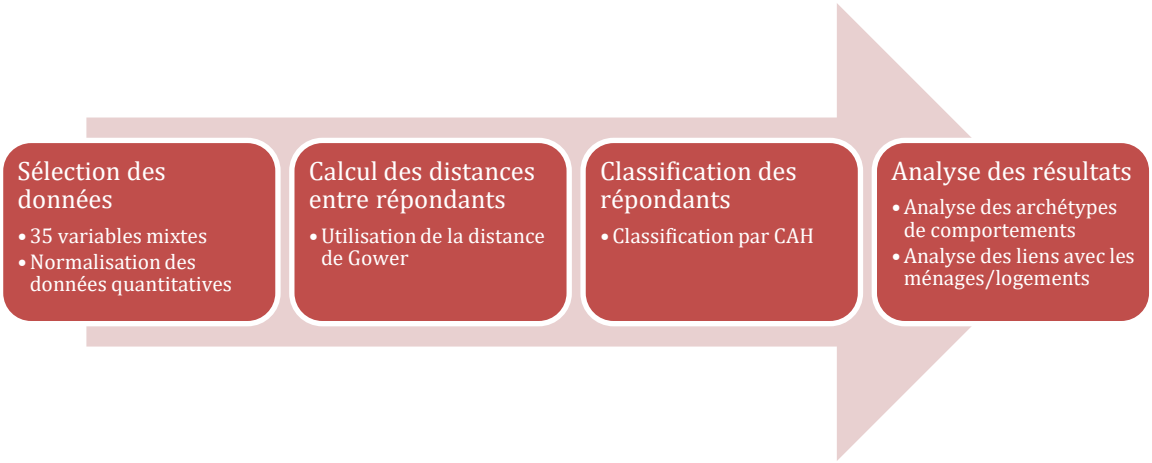
Dans chacune des parties qui suit, on décrit la méthodologie suivie, puis les résultats en mettant en regard les typologies de comportement obtenues avec les variables caractérisant les contextes résidentiels.

Note de vocabulaire : Dans cette partie nous parlerons « d'archétypes de comportement » pour parler des classes construites par segmentation de la base de données composée des variables de comportement. Nous utilisons cette terminologie pour rappeler que les classes construites sont décrites à partir du centre de classe, supposé représenter les assemblages de comportements rattachés à une même classe. La terminologie d'archétype permet cependant de rappeler qu'il ne s'agit que d'un type idéal et que chacun des assemblages est supposé non pas être identique mais plutôt de dériver de ce centre de classe.

2.2 Stratégie S1 : construction d'une typologie par classification des distances de Gower

Méthodologie

La classification des données de comportement suit la méthode suivante. Après avoir calculé la distance de Gower entre les répondants sur les 35 variables de comportement, on effectue une classification ascendante hiérarchique et on analyse les types et les ménages et logements rattachés.



Intérêt de la méthode

Une méthode relativement simple consiste à définir une distance entre les répondants à partir de leurs réponses, puis d'appliquer une méthode de partitionnement qui valorise cette dissimilarité calculée. La distance la plus connue est la distance euclidienne qui est définie pour des vecteurs à valeurs réelles. Celle-ci ne peut être appliquée directement dans notre cas car nous avons des variables qualitatives et quantitatives. La distance proposée par Gower en 1971 offre une alternative en permettant de calculer la similarité entre des individus directement à partir des variables (Gower 1971). Une définition de cette distance est donnée ci-dessous.

Définition de la distance de Gower

La distance entre 2 répondants x_1 et x_2 est définie par la distance de Gower D_g selon :

$$D_g(x_1, x_2) = \frac{1}{p}\sum_{i=1}^{p} s_{12i}$$

Où s_{12i} fait référence à la *similarité* entre x_1 et x_2, et p est le nombre de variables caractérisant les répondants. Pour les variables numériques, s_{12i} est définie par :

$$s_{12i} = \frac{|y_{1i} - y_{2i}|}{\max_{y_i}(|y_{1i} - y_{2i}|)}$$

Avec y_{1i} la valeur prise par la variable i pour le répondant x_1.

Pour les variables qualitatives non ordinales, s_{12i} est définie de la manière suivante, en utilisant la notation δ du symbole de Kronecker[23].

$$s_{12i} = \bar{\delta}_{y_{1i}, y_{2i}}$$

[23] $\delta_{ij} = 1$ si $i = j$ et 0 sinon.

Pour les variables ordinales, les valeurs sont remplacées par l'indice de position correspondant et s_{12i} est calculée de la même manière qu'une variable numérique. Le calcul est effectué sur R à laide de la fonction *daisy* du package *cluster* (v 2.1.4). Cette définition permet ainsi d'ajouter des « contributions » de chacun des variables : la distance absolue pour les variables quantitatives et ordinales et une distance binaire pour les variables nominales. Toutefois, on remarquera que cette distance est par construction très sensible à la sélection des variables.

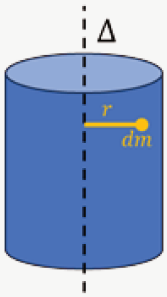
Classification par la méthode de « Classification ascendante hiérarchique » (CAH)

La classification hiérarchique fait partie de la famille des algorithmes de classification automatique (dite aussi « non supervisée » car on n'inclue pas de variable cible dans le travail de segmentation pour guider l'apprentissage). Elle est dite hiérarchique car elle consiste à construire soit de manière descendante soit de manière ascendante une série de classes de données qui vérifie que pour chaque paire de classes, on peut vérifier qu'elles sont soit mutuellement exclusives soit une des deux classes inclue complètement l'autre. La méthode utilisée ici est dite ascendante car les classes sont construites par regroupement des individus. La construction des classes est donc itérative : à partir des n individus à regrouper, on choisit les deux individus dont le « coût » d'agrégation est le plus faible. Ce coût est défini par un indice de dissimilarité. Dans notre travail nous utiliserons l'indice de Ward qui est classiquement utilisé dans la littérature. Cet indice permet de minimiser la perte de l'inertie interclasse dû au regroupement (voir encart sur le calcul de l'inertie ci-dessous).

Inertie totale, inertie inter-classe et inertie intra-classe

Le concept d'inertie utilisé dans les travaux de classification de données est emprunté à la mécanique classique. Dans ce champ, l'inertie désigne la tendance d'un corps à conserver sa vitesse. A titre d'exemple on définit le moment d'inertie d'un corps en rotation autour d'un axe Δ est défini comme :

$$\int_0^R r^2 dm$$



Où r désigne la distance entre l'élément de masse élémentaire dm et l'axe de rotation Δ. R désigne la dimension maximale du corps.

Ainsi, une grande inertie est associée soit à une masse importante soit à une taille importante (ou une combinaison des deux).

Dans le champ de la classification, on qualifie par analogie l'inertie totale d'un jeu de données $\mathcal{T}$ contenant n points, de barycentre G et en utilisant une métrique notée d (par exemple la distance euclidienne) comme étant :

$$\mathcal{I}(\mathcal{T}) = \frac{1}{n}\sum_{M} d(M,G)^2$$

On retrouve ainsi une définition discrète de l'inertie en mécanique où chacun des points M du jeu de données à une masse équivalente et égale à 1. Une illustration d'un jeu de donnée fictif est donnée ci-dessous afin de guider la lecture. On remarquera que la définition dépend fortement de la définition de la distance utilisée. Celle-ci est le plus souvent la norme euclidienne, mais d'autres normes peuvent être utilisées.

En considérant alors une segmentation de l'ensemble de donnée en K classes notées $(\mathcal{T}_k)_{k\in[1,K]}$, chacune ayant un centre d'inertie G_k, on peut définir l'inertie intra-classe comme étant la somme des inerties $\mathcal{I}(\mathcal{T}_k)$:

$$\mathcal{I}_{intraclasse} = \sum_{k} \mathcal{I}(\mathcal{T}_k)$$

Cette inertie intraclasse n'est pas rigoureusement égale à l'inertie totale. D'après le théorème de Huygens il faut y ajouter l'inertie interclasse qui caractérise l'étalement des groupes autour du barycentre G. La décomposition de l'inertie s'écrit donc :

$$\mathcal{I}_{totale} = \mathcal{I}_{intraclasse} + \mathcal{I}_{interclasse}$$

Avec :

$$\mathcal{I}_{interclasse} = \frac{1}{n}\sum_{k} n_k \, d(G_k,G)^2$$

On comprend intuitivement que lorsqu'on augmente le nombre de classes on diminue l'inertie intra-classe et on augmente l'inertie interclasse. Dans le cas de la CAH, on débute le calcul avec une inertie interclasse maximale (il y a autant de clusters que de points $\mathcal{I}_{intraclasse} = 0$). Le regroupement fait baisser cette inertie interclasse et la méthode de Ward permet de minimiser cette perte.

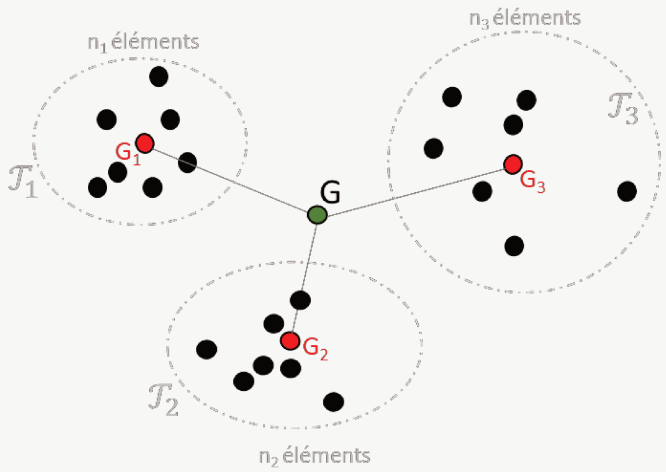


Figure 21: Exemple graphique d'un jeu de données en 2 dimension segmentée 3 classes. Le barycentre G du jeu de donnée et les barycentres des classes G1, G2 et G3 sont représentés. Source : Auteur

Le regroupement itératif permet de construire un dendrogramme, qui est un graphique qui peut être vu comme un « arbre » dont les « feuilles » sont les individus (Figure 22). Chaque nœud marque le regroupement de deux « branches » et la hauteur de ce nœud est égale à la dissimilarité entre ces deux groupes : deux groupes très différents seront ainsi « noués » tardivement dans le processus et donc auront un nœud assez haut dans ce graphe. Au contraire, les individus qui ont une faible distance seront regroupés rapidement et leur nœud sera assez bas.

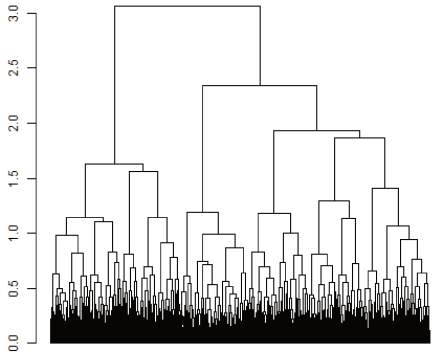


Figure 22 : Exemple de dendrogramme. Source : Auteur

Analyse des résultats

Le calcul de l'inertie en fonction du nombre de clusters montre qu'il est intéressant de retenir une classification à 7 classes. En revenant aux variables initiales on peut alors qualifier les tendances en termes d'équipement, de chauffage, de régulation et d'occupation du logement pour chacun des types. On associe dans la description des archétypes les caractéristiques des ménages et des logements (les situations d'habitation) des profils typiques ainsi que la distribution des consommations d'énergie finale pour chacun des groupes (donnée à la Figure 23). Cette description permet alors de mieux comprendre les logiques de consommation des clusters.

Tableau 11 : Synthèse des archétypes de comportement obtenus par classification selon la méthode S1. Les types sont décrits à partir de l'analyse de la distributions des variables initiales dans chacune des classes. Source : Auteur.

Archétype	Description des comportements typiques et des contextes résidentiels associés.
GOW_1 « Abri économique »	Ce cluster regroupe des ménages composés de personnes jeunes et actives, plutôt avec des revenus bas, et vivant seules comme locataire ou propriétaire accédant d'un logement collectif. En raison de leur activité, ces ménages sont peu présents et vraisemblablement en raison des revenus limités, ces ménages indiquent en moyenne avoir peu d'équipements domestiques par rapport à la moyenne et mettre en place des stratégies de régulation du chauffage.
GOW_2 « Maison économique »	Le cluster 2 est proche du cluster 1 en termes de comportements et de ménages. Il diffère par le niveau des comportements de restriction qui y sont bien plus importants. Aussi, le cluster est composé en proportions par plus de couples et de personnes seules avec des enfants.
GOW_3 « Maison pratique »	Ce cluster se caractérise par un niveau d'équipement et d'usage dans tous les domaines très important. Seul le chauffage fait l'objet de gestes de régulation. En termes de ménages et de logement, on retrouve dans ce cluster plutôt des couples d'actifs avec plusieurs enfants, avec des revenus moyens voire élevés, vivant plutôt dans un logement individuel qu'il possède. Avec le cluster 6, ce groupe présente en moyenne les consommations en énergie finale les plus importantes.
GOW_4 « Cocon frugal »	Ménages composé d'une personne seule de plus de 70 ans, aux revenus faibles, plutôt issue de milieu ouvrier et étant propriétaire mais ayant une probabilité importante d'être locataire d'un logement collectif ancien. En termes de comportements, ces ménages couplent une occupation très importante au logement avec un niveau d'équipement très faible et des gestes de régulation importants, aboutissant *in fine* à une consommation en énergie finale tout juste inférieure à la moyenne.
GOW_5	Le cluster 5 est relativement difficile à interpréter car il diffère des autres uniquement par une faible équipement en hygiène et des gestes de régulation du chauffage. Par ailleurs les ménages le composant sont caractérisés principalement par le fait d'avoir des revenus plus bas que la moyenne.
GOW_6 « Cocon loisirs »	Ce cluster comprend des ménages qui sont plutôt des couples avec enfants, où plus âgés et dont les enfants sont partis du domicile. Ces ménages ont des revenus très élevés et ont de fortes chances d'être à la retraite. Ils sont propriétaires de leur logement individuel ou collectif, plutôt ancien et qui possède une surface en moyenne de plus de 100 m². Les ménages de ce groupe couplent généralement des comportements énergivores dans tous les domaines de consommation, aboutissant aux consommations en énergie finale les plus importantes en comparaison des autres clusters.
GOW_7 « Maison familiale »	Le cluster 7 comprend des ménages composé de couples autour de 40, plutôt de classe moyennes ou populaires ils ont un ou deux enfants et sont actifs. Ils occupent un logement souvent social ou sont propriétaires. Leur logement a une probabilité importante d'être plus récent ou rénové. Il est intéressant de remarquer qu'en comparaison du cluster 6 avec qui les ménages partagent un mode de vie énergivore, le cluster 7 présente une distribution des consommations plus faible. Cela peut être expliqué notamment par une présence plus faible, une isolation thermique plus performante.

Cette première classification est intéressante dans le sens où elle permet de dégager un premier aperçu des liens les plus récurrents entre les variables de comportement. On observe en particulier une association forte entre les niveaux d'équipement et la position du ménage dans son cycle de vie, éventuellement modulé par le niveau de revenu et l'âge. Cette première classification repose sur un critère simple, dans le sens où elle ne tient pas compte de la redondance de l'information contenue dans des variables corrélées et qu'elle donne un poids égal à chacune des variables initiales. La force de ce premier calcul est de montrer que même avec un critère simple, des associations entre occupation, équipement comportements de régulations peuvent être calculés et que ceux-ci sont associables à des types de ménages singuliers. Aussi, on observe que la classification des variables de comportement permet de distinguer des groupes de ménages et de logements qui diffèrent en termes de consommation en énergie finale totale mais sont plutôt proches en termes de consommation par mètre carré et par personne (Figure 23).

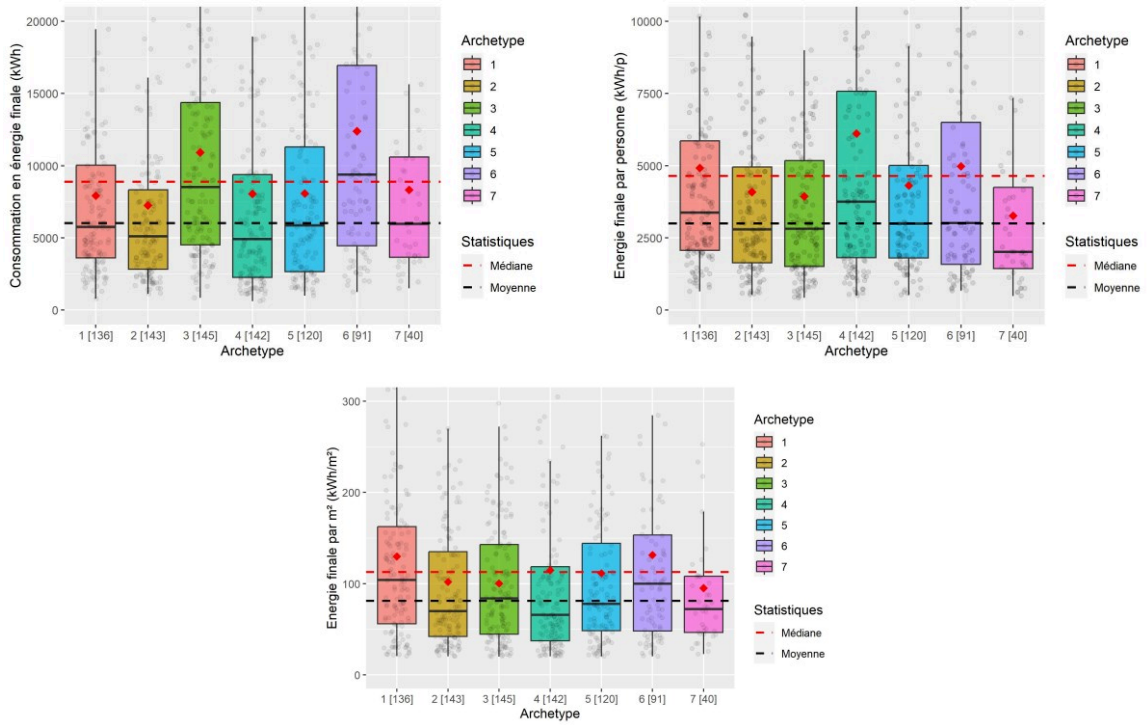


Figure 23 : Distribution des consommations en énergie finale pour chacun des archétypes calculés par la stratégie S1. On différencie 3 indicateurs de consommation : la consommation totale, par personne, et par mètre carré du logement. Source : Auteur après calculs sur la base ENERGIHAB.

La principale faiblesse de cette méthode est sa dépendance aux variables sélectionnées et à leur corrélation. Nous proposons dans la stratégie S2 d'ajouter une étape de calcul qui consiste à calculer un espace synthétique de représentation des variables initiales qui permettrait à la fois de mieux cerner les « dimensions » comportementales qui permettent de caractériser les modes de vies résidentiels et de fiabiliser le calcul des archétypes de comportement.

2.3 Stratégie S2 : construction d'une typologie par analyse factorielle de données mixtes (AFDM)

Une seconde approche consiste à construire un espace synthétique à l'aide d'une analyse factorielle puis effectuer une classification sur les scores des ménages dans cet espace à l'aide d'une CAH. Dans cette partie nous décrivons l'intérêt de la stratégie, les étapes de calcul et les hypothèses prises, avant de présenter les résultats de la classification.

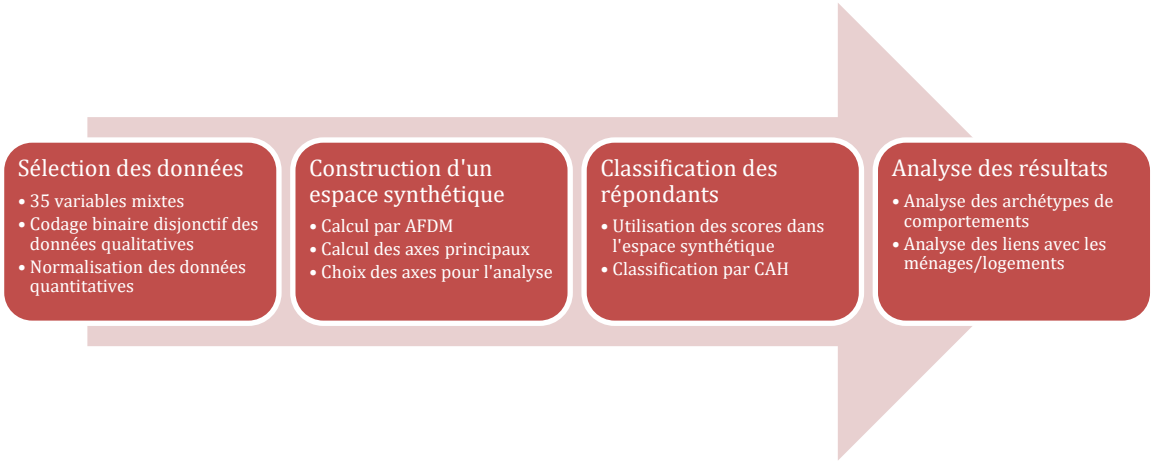


Figure 24 : Méthodologie suivie pour construire des archétypes de comportements (stratégie S1 : AFDM + CAH). Source : Auteur.

Le choix de la méthode de l'analyse factorielle de données mixtes

En raison de leur complexité, les comportements résidentiels sont susceptibles d'être décrits dans des espaces de grande dimension avec des variables quantitatives et qualitatives. Comme nous l'avons vu dans la méthode précédente, le calcul d'une distance dans ces espaces et la réalisation d'une classification est instable et dépend de l'échantillon et des variables sélectionnées. Par ailleurs, nous avons vu que l'interprétation des résultats n'est pas aisée. Une autre approche peut être de construire une espace de plus petite dimension (quelques variables plutôt que quelques dizaines) et de réaliser ensuite une classification dans cet espace. Plusieurs méthodes existent et l'analyse factorielle offre une série d'algorithmes dont l'analyse en composantes principales et l'analyse en composantes multiples pour les variables quantitatives et qualitatives. Ces deux algorithmes sont très utilisés dans la littérature et leur manipulation est largement facilitée par la disponibilité de nombreux outils (librairies et logiciels) en libre accès sur internet. Le critère utilisé pour construire ces espaces est l'inertie, c'est-à-dire la dispersion des individus. Les axes de l'espace sont construits de manière itérative. A chaque étape un axe est construit comme combinaison linéaire des variables initiales, orthogonal à l'axe précédent et avec comme objectif de maximiser l'inertie projetée des individus sur cet axe (Figure 25).

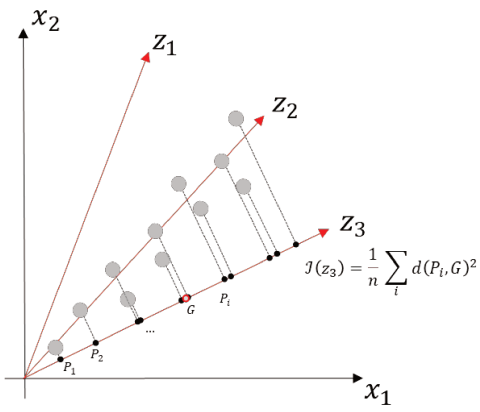


Figure 25 : Illustration d'un nuage de points de coordonnées (x_1, x_2). La construction d'un axe factoriel implique de rechercher les coefficients a et b tels que la variable construite $\mathbf{z_i} = a.\mathbf{x_1} + b.\mathbf{x_2}$ maximise l'inertie $\mathcal{I}(\mathbf{z_i})$ des points projetés sur $\mathbf{z_i}$. G désigne le barycentre sur z_3 des projections des points sur $\mathbf{z_3}$. Dans cette figure c'est la variable z_2 qui réalise cette condition. Source : Auteur.

Pour manipuler des données mixtes, la plupart des travaux ayant mobilisé une analyse factorielle ont créé des variables catégorielles à partir de variables quantitatives par seuillage pour former un ensemble homogène de variables catégorielles. Cette approche présente l'inconvénient de perdre une quantité importante d'informations contenues dans les variables quantitatives et d'augmenter la dimension du problème. Plus récemment, une approche alternative en analyse factorielle a permis de traiter ces ensembles de données simultanément. Cette méthode est connue sous le nom d'analyse factorielle de données mixtes (AFDM). Une présentation exhaustive ainsi qu'une revue de littérature sur cet algorithme développée depuis les années 1980 peut être trouvée dans l'article de Jérôme Pagès (Pagès 2002). De même que pour une Analyse en composante principale (ACP – utilisée pour les données quantitatives) ou que pour une Analyse en Composantes Multiples (ACM- pour les données qualitatives) cet algorithme permet de construire des facteurs synthétiques orthogonaux (appelées Composante Principales), par combinaison linéaire des variables de départ. Ces facteurs fournissent ainsi des scores synthétiques qui sont utilisés pour effectuer une classification hiérarchique agglomérative (CAH). L'AFDM constitue ainsi une opportunité de dépassement des travaux de segmentation par ACM ou ACP.

Il est important toutefois de noter que ainsi que De Soete et Carroll (1994) l'ont écrit, l'espace des facteurs principaux orthogonaux ne permet pas toujours une représentation adéquate de l'information taxonomique, limitant de fait la capacité de cette famille d'algorithmes à fournir un espace pertinent pour la construction d'une typologie. Cet effet n'est toutefois pas systématique et le cadre de l'analyse factorielle permet de fournir *a minima* une bonne manière de s'approprier un ensemble de données, et souvent de fournir une segmentation intéressante pour de nombreux problèmes.

Introduction à l'AFDM

L'analyse factorielle est une famille d'algorithmes qui permettent d'analyser les jeux de données. Ils permettent une approche exploratoire et compréhensive des données à disposition. La méthode générale consiste à construire des axes principaux comme des combinaisons linéaires des variables initiales de telles manières que ces axes résument la complexité d'un tableau avec les oppositions les plus importantes présentes entre les individus. En termes mathématiques, on dit que les premiers axes principaux résument une part de l'inertie totale plus importante que les axes suivants. En général, on se contente de s'intéresser aux 2 ou 3 premiers axes de l'analyse ce qui permet alors de tracer les individus dans un nouvel espace en deux ou en 3 dimensions. L'observation de la dispersion entre les individus sur un graphique permise par l'analyse factorielle permet alors de mieux saisir la répartition des données.

Dans le cas où les données sont uniquement catégorielles (resp. quantitatives), on parle d'analyse en composante multiples – ACM (resp. analyse en composantes principales). Lorsqu'une table de données contient des données quantitatives, une pratique courante est de discrétiser les variables puis de procéder à une ACM. Cette pratique peut en revanche être problématique car le choix des seuils influe sur les résultats. Une alternative consiste à réaliser une analyse factorielle de données mixtes (AFDM). Cette méthode consiste en fait à réaliser le codage binaire des variables qualitatives, puis à pondérer ce codage selon les fréquences des modalités et en fin à réaliser une ACP normée sur ce tableau (Pagès 2002).

Analyse des résultats

Interprétation des axes factoriels

Le calcul de l'AFDM permet de générer plusieurs objets à analyser. En premier lieu, la variance résumée par chacun des axes factoriels permet d'avoir un premier a priori de la « complexité » du set de données. En effet, si un axe permet de résumer 90% de la variance, cela signifie que le tableau de données peut être résumé par une seule variable quantitative en perdant uniquement 10% de la variabilité initiale. Dans notre cas, on observe sur la Figure 26, la variance expliquée par chacun des axes. L'axe 1 résume ici environ 10% de l'information, les 10 premiers axes en résument 41% et 42 axes sont nécessaires pour résumer 100% de l'information. On constate donc que la synthèse des variables de comportement nécessite de garder un nombre important de facteurs principaux si l'on souhaite ne pas avoir trop de perte d'information. Un tableau synthétisant la contribution et la projection de chacune des variables sur les axes principaux est donné en annexe (p. 230).

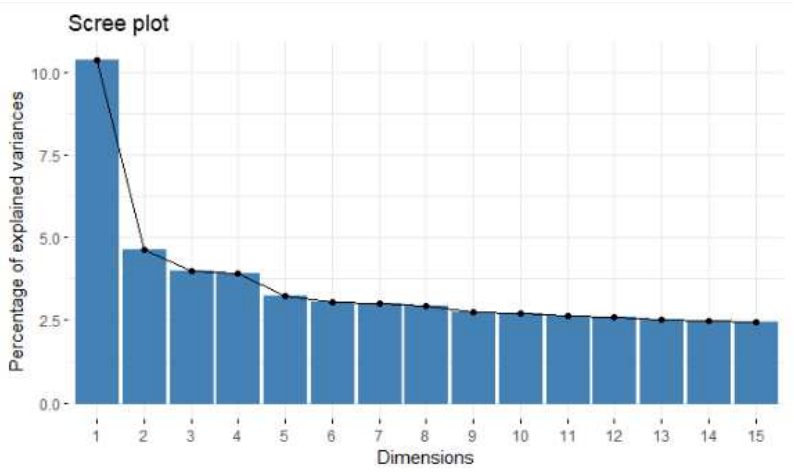


Figure 26 : Pourcentage de variance expliquée par chacun des axes principaux. Source : Auteur, calculs effectués sur la base ENERGIHAB.

Dans un second temps on peut s'intéresser à la contribution des variables à la construction des axes factoriels et à la projection de celles-ci sur les axes principaux. Tandis que la première information permet d'identifier les variables qui ont contribué à construire l'axe principal étudié, la seconde permet de qualifier et d'interpréter la variable quantitative (l'axe principal) construite. Pour cela, nous observons pour les variables initiales quantitatives leur coefficient de corrélation avec l'axe principal. Pour les variables qualitatives dont la contribution à l'axe principal est suffisamment importante, nous observons la position des modalités sur l'axe principal. Nous pouvons ainsi donner du sens aux 6 premiers axes principaux (AP) que nous choisissons de retenir pour la suite de ce travail :

- **AP1** : Niveau d'équipement
- **AP2** : Présence à la maison
- **AP3** : Besoin en chauffage
- **AP4** : Type d'éclairage
- **AP5** : Demande en loisirs
- **AP6** : Pratiques de régulation

La Figure 27A représente les lignes de la base de données dans le repères des deux premiers axes principaux. On observe que les points situés au centre ont une qualité de projection faible ($cos^2<0,1$) au contraire des points situés à la périphérie ($cos^2>0,4$). Cela dénote que les ménages au centre ne sont pas différenciés par les variables qui ont servi à la construction des axes AP1 et AP2. Cet exemple visuel permet d'illustrer comment il est possible de caractériser les archétypes de comportement construits par CAH.

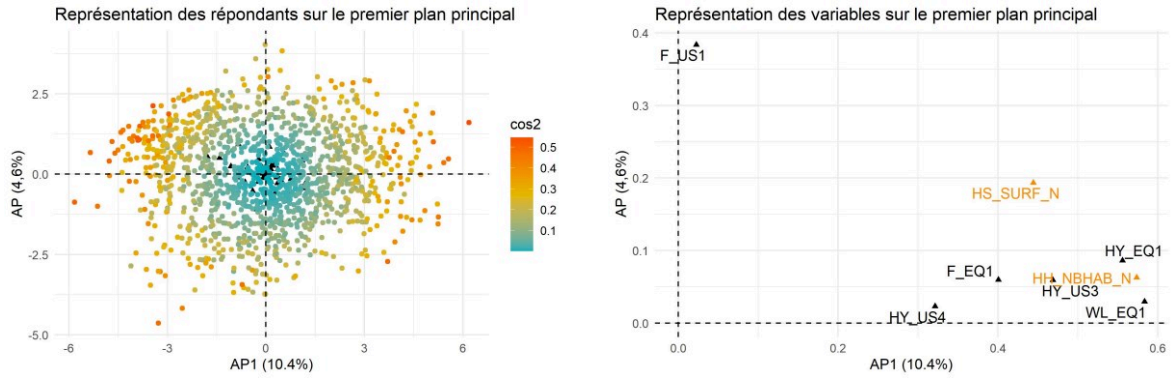


Figure 27 : Projection des répondants et des variables de comportements dans le repère des deux premières composantes principales. A gauche, les couleurs indiquent la qualité de la représentation de l'individu sur ce plan factoriel. Un cos² trop faible indique une projection peu significative. A droite, seules sont représentées les variables avec une qualité de projection suffisante sur ce plan (cos²>0,2). Source : Auteur à partir de calcul sur les données ENERGIHAB.

Analyse des archétypes

De la même manière que dans la partie précédente, nous proposons d'articuler la description des archétypes de comportement calculés avec les variables caractérisant les logements, les ménages et les consommations d'énergie associées. Une synthèse de la caractérisation des centres des classes calculées par CAH est donnée dans le Tableau 12.

Tableau 12: Présentation des caractéristiques moyennes en termes de comportements pour chaque archétype. L'interprétation qualitative est obtenue par analyse de la position des centres de classes dans l'espace des composantes principal. Source : Auteur après calculs sur la base ENERGIHAB.

Archétype	AFDM_1	AFDM_2	AFDM_3	AFDM_4	AFDM_5	AFDM_6	AFDM_7
Nom de l'archétype Axes principaux (interprétation)	Cocon frugal	Maison sobre	Maison chaude	Maison verte	Maison loisirs	Maison économe	Maison confort
AP1 : Niveau d'équipement	Très faible	Très faible	Moyen	Moyen	Elevé	Elevé	Très élevé
AP2 : Présence à la maison	Plutôt présent	Très peu présent	Moyenne	Peu présent	Très présent	Peu présent	Moyenne
AP3 : Besoin en chauffage	Moyen	Faible	Très élevé	Moyen	Faible	Très faible	Elevé
AP4 : Type d'éclairage	LED fréquentes	LED rares	LED rares	LED systématiques	LED fréquentes	LED rares	LED fréquentes
AP5 : Demande en loisirs	Faible	Faible	Très élevé	Elevé	Elevé	Très faible	Très faible
AP6 : Pratiques de régulation	Importantes	Faibles	Importantes	Moyennes	Faibles	Très importantes	Moyennes
Taille de l'échantillon [Pour lesquels on dispose de données de consommation d'énergie]	**262 [154]**	**125 [82]**	**179 [112]**	**197 [127]**	**172 [101]**	**216 [127]**	**212 [114]**

On peut décrire archétype par archétype les profils typiques des ménages et des logements, en relation avec les distributions des consommations d'énergie finale (Figure 28).

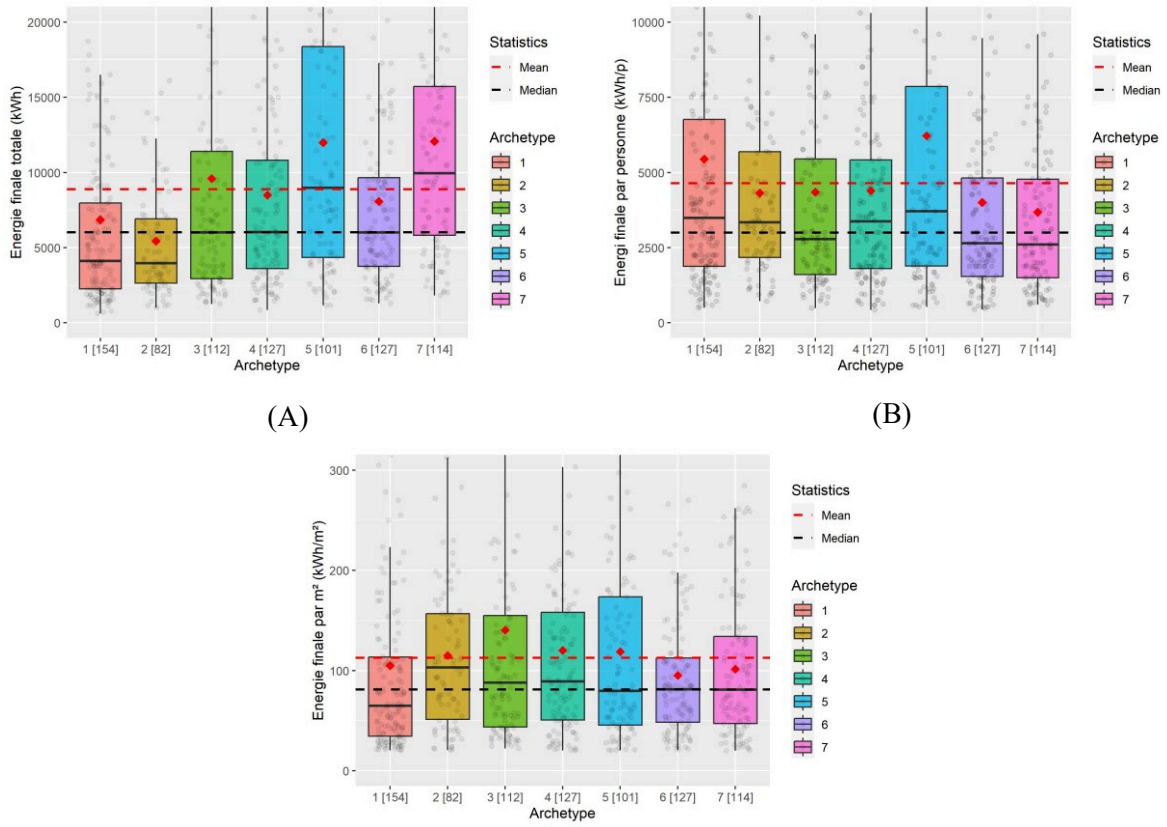


Figure 28 : Tracé des distributions des consommations en énergie finale totale (A), par personne (B) et par mètre carré(C) pour chacun des archétypes de comportements calculés par AFDM. Source : Auteur après calculs sur la base ENERGIHAB

- **AFDM_1 « Cocon frugal » :** les ménages composant cet archétype sont plutôt des personnes âgées seules et aux revenus faibles. Ils occupent souvent en tant que locataire un appartement plutôt ancien mais ayant en moyenne une surface plus importante que dans l'échantillon. En termes de comportement, on observe une forte présence, un équipement plutôt faible et des gestes de régulation du chauffage plutôt importants. En termes de consommation, ces ménages ont une consommation totale plus faible que la moyenne.
- **AFDM_2 « Maison sobre »** : cet archétype regroupe des ménages composé plutôt de personnes seules plutôt jeunes et actives. Les ménages ont des revenus faibles et sont locataires dans le parc privé d'un logement collectif, généralement entre 30 et 50m². Ces ménages associent dans leur quotidien une faible présence au logement, un faible équipement et une demande en chauffage plutôt faible. La logique économique décrite aboutit à une consommation totale très faible en comparaison des autres archétypes.
- **AFDM_3 « Maison chaude »** : cet archétype se distingue des autres par la demande en chauffage particulièrement élevée, associée à des pratiques de régulation importantes. Les autres comportements sont eux dans la moyenne de l'échantillon. La logique de confort thermique qui définit cet archétype est à mettre en regard avec le fait qu'il est difficile d'identifier des profils typiques de ménages et de logement à cet archétype. Cette remarque permet alors de faire les

hypothèses suivantes : (1) les pratiques de chauffage sont peu liées aux autres comportements, et (2) les variables explicatives de ces pratiques ne sont pas inclues dans la définition des situations d'habitation que nous avons retenue.

- **AFDM_4 « Maison verte »** : les ménages rattachés à cet archétype de comportement ont des niveaux d'équipement moyens hormis en termes de loisirs, sont globalement peu présents au logement. Les ménages sont plutôt de jeunes couples avec un enfant. En début de cycle de vie, ils ont plutôt peu de revenus et occupent leur logement en tant que locataires dans le parc public ou le parc privé.
- **AFDM_5 « Maison loisirs »** : cet archétype de comportement se caractérise par un fort niveau d'équipement, une forte présence au logement et peu de gestes régulation. Les ménages associés sont plutôt des couples de personnes âgées de plus de 70 ans, retraitées et sans enfants au domicile, et propriétaire de leur logement.
- **AFDM_6 « Maison économe »** : cet archétype de comportement se caractérise par un fort niveau d'équipement et d'usage mais aussi avec une demande en chauffage très faible ainsi que des gestes de régulations nombreux. Aussi, on note une faible présence au logement. Les ménages composant cet archétype sont principalement de jeunes couples avec enfants, propriétaires de leurs logement de plus de 75 m², souvent plus récent que la moyenne de l'échantillon. *In fine* ces ménages présentent une consommation en énergie finale totale relativement faible en comparaison avec l'échantillon total.
- **AFDM_7 « Maison confort »** : ce cluster présente des comportements proches du cluster AFDM_6 à l'exception de la demande en chauffage qui est là très importante, en association avec des pratiques de régulation plus faibles. Dans ce cluster, on observe là aussi des ménages composés de couples avec enfants aux revenus élevés et vivant dans des maisons individuelles. Ces ménages présentent quant à eux des consommation énergétique totale élevées en moyenne.

La description des archétypes obtenue par cette stratégie S2 aboutit sur une classification proche de celle obtenue dans la partie précédente. Elle montre également une association entre occupation, équipement, et gestes de régulation en relation avec les situations d'habitation, où les variables caractérisant le revenu et la composition du ménage occupent une place importante.

Pour faciliter la visualisation de la distribution des archétypes selon ces deux variables nous avons représenté les proportions des archétypes en fonctions des tranches d'âge et en séparant les ménages en fonction de leurs revenus relativement à la médiane (Figure 29). La lecture des deux figures permet de visualiser des distributions très différentes selon l'âge de la PR et le niveau de revenu du ménage.

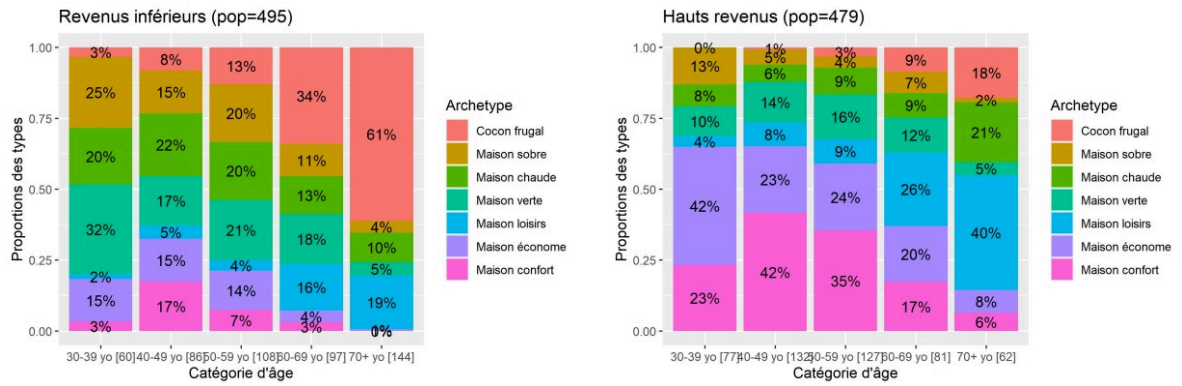


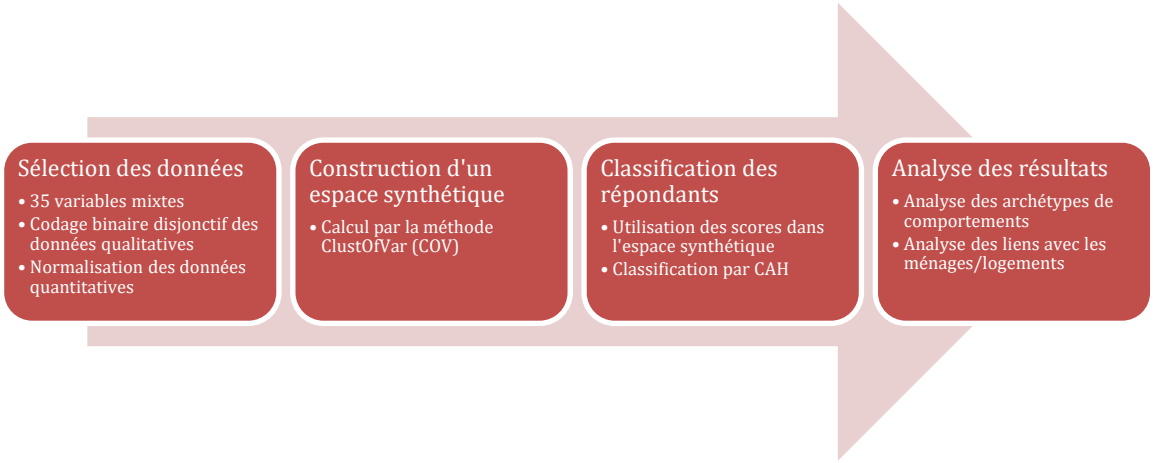
Figure 29 : Tracé des proportions des archétypes par tranche d'âge et selon le niveau de revenus. Les ménages dont la PR est âgée de moins de 30 ans ne sont pas représentés en raison d'effectifs trop faibles. Source : Auteur d'après calculs sur la base ENERGIHAB.

La stratégie S2 impliquant la réduction de dimension par analyse factorielle de données mixtes a facilité la classification des données de comportement : en construisant un espace de dimension réduite il devient d'une part plus facile d'interpréter les archétypes construits et d'autres part il est possible d'identifier des dimensions « orthogonales » c'est-à-dire indépendantes permettant de décrire les comportements domestiques. La principale limite de cette méthode réside toutefois dans l'interprétation de ces axes qui est difficile qui fragilise l'analyse et l'exploitation. Une alternative pourrait être d'exploiter la corrélation des variables de comportement pour construire des dimensions comportementales non orthogonales mais pouvant être plus facilement interprétées.

2.4 Stratégie S3 : construction d'une typologie par regroupement de variables mixtes (ClustOfVar)

La méthode « ClustOfVar » (abrégée en COV) que nous proposons dans cette 3e stratégie de modélisation consiste à réaliser une classification des variables de comportement en utilisant un critère de corrélation, puis de réaliser une CAH dans l'espace synthétique construit. Le détail de la construction des variables a été donné dans la partie II.1.2.2 .

Méthodologie



Les clusters de ménages sont construits par classification ascendante hiérarchique (CAH) à partir des 9 variables synthétiques calculées dans la partie (II.1.2.2) après centrage et réduction. La distance entre deux ménages est calculée comme la distance euclidienne entre leurs coordonnées normalisées. Le critère d'agrégation pour la classification hiérarchique est le critère de Ward (Saporta, 2006) .

Le choix du nombre K_2 de clusters de ménages a été effectué de manière itérative pour obtenir des clusters homogènes et interprétables. Dans un premier temps, nous avons tracé les gains d'inertie intra-clusters en fonction du niveau de partitionnement. Pour satisfaire le critère d'homogénéité, nous avons choisi le nombre K_2 de clusters pour lequel la diminution de l'inertie intra-cluster entre K_2 - 1 et K_2 clusters était beaucoup plus importante que celle entre K_2 et K_2 + 1 clusters. Pour répondre au critère d'interprétabilité, nous avons observé les distributions des comportements, des caractéristiques des logements et des profils des ménages représentés dans les K_2 clusters.

L'inertie interclasse est tracée en fonction du nombre de clusters (Figure 30). Cinq niveaux de K_2 ont été considérés comme particulièrement intéressants pour le clustering (K_2=2, K_2=4, K_2=7, K_2=10, K_2=13). Le choix de K_2=13 aurait pu produire des clusters de comportement très précis, mais les clusters obtenus n'étaient pas interprétables. Comme compromis entre l'homogénéité des clusters et la parcimonie, nous avons d'abord choisi K_2= 10, mais un cluster comprenant environ 100 ménages n'était expliqué que par une seule VS, ce qui n'était pas utile pour construire des clusters explicables. Finalement, nous avons choisi K_2=7.

L'analyse des résultats du clustering s'est faite en trois étapes. Tout d'abord, les coordonnées des barycentres des archétypes ont été calculées. Un test de Student a été effectué sur les coordonnées des barycentres des archétypes pour identifier les VS qui caractérisent le mieux les clusters construits. En utilisant la définition de la VS les barycentres peuvent être décrits qualitativement. L'homogénéité des clusters construits est examinée en calculant la distance moyenne des ménages par rapport au barycentre de leur archétype.

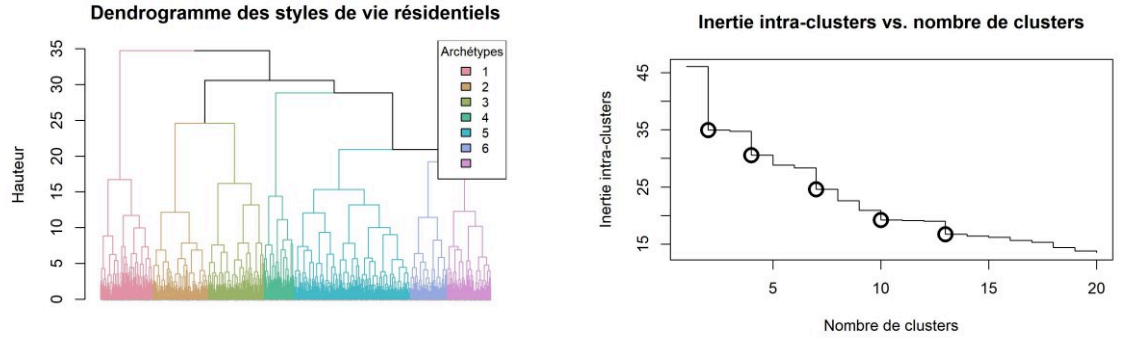


Figure 30 : Dendrogramme des styles de vie résidentiels (gauche) et évolution de l'inertie interclasse en fonction du nombre de classes (droite). Source : Auteur d'après calculs sur la base ENERGIHAB.

Ensuite, un test statistique d'indépendance a été réalisé pour vérifier si les archétypes de comportement et 20 variables caractérisant les ménages, les logements et les utilisations de l'énergie sont indépendants.

Pour les variables qualitatives, le test effectué est un test du Khi-deux. Pour les variables quantitatives, la valeur p a été calculée par ANOVA et correspond au risque de rejeter l'hypothèse d'égalité des moyennes entre les archétypes. En raison de données insuffisantes, seuls les ménages qui consomment du gaz et/ou de l'électricité ont été considérés pour cette étape. Les intensités énergétiques ont été calculées en tant qu'énergie finale (en kWh_{FE}) car elles visent à représenter la consommation d'énergie en tant que service. Il faut souligner que les données de consommation sont autodéclarées et basées sur les factures et souffrent donc d'imprécision. Une étude réalisée en 1984 par Warriner (Warriner, McDougall,et Claxton 1984) a estimé que cette erreur pouvait être comprise entre 10,5 % et 29,3 % selon le type d'énergie et l'utilisation des reçus des ménages. Seule la consommation d'énergie comprise entre 20 kWh_{FE} /m² et 1000 kWh_{FE} /m² a été considérée comme correcte, correspondant à 575 des 1363 ménages. La répartition des ménages et des logements était similaire dans ce sous-ensemble, ce qui a permis de poursuivre l'analyse.

Enfin, des profils types de ménages et de logements ont été construits pour chaque archétype. Pour ce faire, nous avons effectué une analyse factorielle couplée à une classification ascendante hiérarchique sur chaque ensemble de ménages associés aux archétypes. Ce calcul nous permet de décrire les liens entre les ménages, les logements et l'archétype et de faire des hypothèses sur la logique de consommation sous-jacente.

Analyse des archétypes construits

Pour caractériser l'homogénéité des clusters, nous avons calculé la moyenne et l'écart-type des distributions de distance au barycentre de chaque type. Les scores des barycentres sur les VS de chaque cluster sont donnés en annexe (voir p. 230). Les clusters obtenus sont assez homogènes. Les distances moyennes au barycentre sont comprises entre 1,7 et 2,1 sauf pour le cluster 5 (moyenne 2,7). Les clusters 4, 5 et 7 ont une dispersion plus importante (écart-type de la distribution des distances de 2, 2,7 et 2,1 respectivement) que les autres clusters (écart-type autour de 1,7), et peuvent donc être interprétés comme ayant une plus faible homogénéité. Les distances moyennes au barycentre peuvent être interprétées en termes de comportement en observant la position des modalités sur la VS. Ce faisant, nous avons constaté que chaque ménage se distingue en moyenne du barycentre, considéré comme représentatif du type, par la modulation d'un comportement dans la grille des VS. Dans la suite de l'article, nous nous concentrerons sur le comportement moyen des archétypes, c'est-à-dire sur l'ensemble des comportements qu'adopterait un ménage dont les coordonnées sur la VS sont celles du barycentre des ménages classés dans le type. Il convient toutefois de rappeler que les ménages regroupés au sein des archétypes adoptent des comportements qui peuvent être considérés comme des modulations de ce comportement moyen.

Pour faciliter l'interprétation des archétypes comportementaux, les coordonnées du barycentre de chaque classe ont été traduites en termes de comportement. L'interprétation des VS a été réalisée en exploitant

les coordonnées des modalités des variables qualitatives sur les VS et les corrélations des variables quantitatives avec les VS. Les résultats sont présentés dans le Tableau 13. Seules les coordonnées non nulles (au risque de 10%) sont interprétées. Enfin, un nom a été associé à chaque archétype de comportement en fonction des caractéristiques saillantes (telles que le niveau de température, le niveau d'équipement, la présence dans le logement). Un code couleur a été utilisé pour mettre en évidence la tendance d'un comportement à diminuer (vert) ou à augmenter (rouge) la consommation finale d'énergie.

Tableau 13 : Synthèse des archétypes de comportements construits à l'aide de la stratégie de modélisation S3, basée sur la méthode ClustOfVar. Le code couleur permet d'identifier des comportements moyens a priori énergivore (rouge) ou économes en énergie (vert). Source : Auteur, après calculs sur la base ENERGIHAB.

Archétype Nom de l'archétype	COV_1 Maison pratique	COV_2 Maison économe	COV_3 Maison cocon	COV_4 Maison chaude	COV_5 En quête de confort	COV_6 Maison frugale	COV_7 Maison minimale
VS1 : Equipement alimentaire	Élevé	Bas	Très élevé	Moyen	Bas	Très bas	Moyen
VS2 : Présence au logement	Plutôt basse	Plutôt basse	Très élevée	Très basse	Plutôt basse	Élevée	Très élevée
VS3: Equipment en hygiène	Élevé	Bas	Très élevé	Élevé	Moyen	Très bas	Bas
VS4 : Pratiques de restriction	Basses	Moyennes	Plutôt élevées	Basses	Très élevées	Moyennes	Basses
VS5 : Besoin en chauffage	Plutôt élevé	Très bas	Plutôt bas	Très élevé	Moyen	Moyen	Moyen
VS6 : Comportements de régulation du chauffage	Importants	Très importants	Peu importants	Peu importants	Moyen	Peu importants	Peu importants
VS7 : Demande en loisirs numériques	Très élevée	Basse	Très élevée	Élevée	Moyenne	Très basse	Basse
VS8 : Gestes verts	Plutôt importants	Importants	Importants	Plutôt importants	Moyen	Moyen	Peu importants
VS9 : Equipements en luminaires	LED tout le temps	LED rarement	LED parfois	LED parfois	LED souvent	LED souvent	LED parfois
Nombre de ménages	227	349	251	192	53	198	93

Ces ensembles de comportements peuvent être interprétés à partir des associations récurrentes entre les variables. Des associations récurrentes entre les niveaux d'équipement (SV1 et SV3) peuvent alors être observées pour la plupart des archétypes. On peut alors s'intéresser à des ensemble de comportements énergivores, combinant des niveaux d'équipement et des besoins de chauffage élevés (archétypes 1 et 4) ou, au contraire, à des comportements frugaux (archétypes 6 et 7). Cependant, l'analyse des associations comportementales entre la présence au domicile, l'équipement et le chauffage n'est pas simple et l'ajout de connaissances sur la consommation d'énergie, les ménages et les habitations permet d'affiner leur analyse.

Analyse du lien entre les typologies, les contextes résidentiels et les consommations d'énergie

Les calculs de test d'indépendance montrent un lien étroit entre les archétypes et les variables décrivant les ménages et les logements. Certaines variables ne semblent cependant pas liées. La classe socioprofessionnelle, la surface par personne et le type d'énergie ne semblent pas liés aux archétypes de comportement. Par ailleurs, il faut noter que la consommation d'énergie finale est fortement liée aux

archétypes, alors que la consommation par m² et par personne ne le sont pas (voir Figure 31). Ce résultat semble montrer que les archétypes de comportement se réfèrent essentiellement aux structures des ménages et des logements de différentes tailles, qui sont des facteurs clés de la consommation d'énergie. En corollaire, compte tenu de l'absence de lien significatif entre les archétypes et la consommation par m² et par personne, on peut également affirmer que la différence éventuelle de consommation entre des archétypes plus ou moins énergivores est inférieure à l'erreur de mesure (entre 10 et 30%).

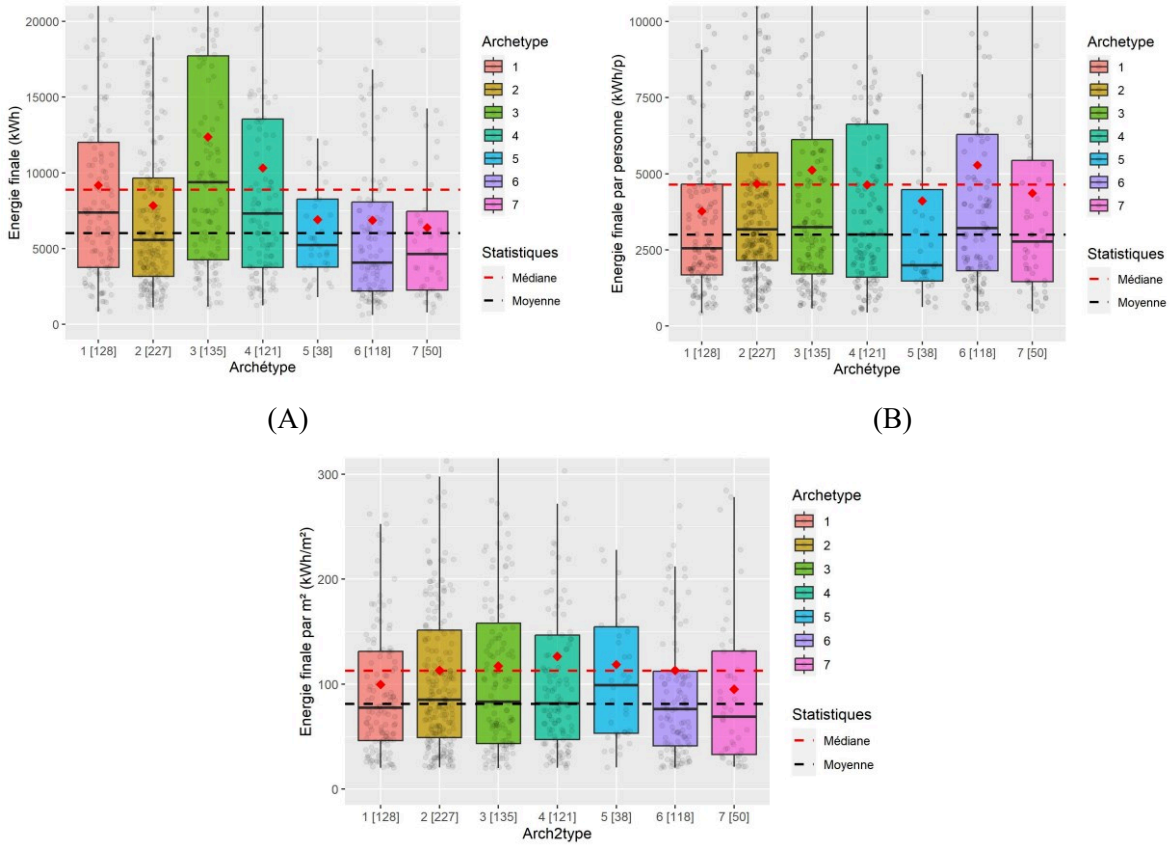


Figure 31 : Distributions des consommations d'énergie finale totale (A), par personne (B), et par m² de surface du logement (C). Source : Auteur, d'après calculs sur la base ENERGIHAB.

Finalement, en faisant le lien entre les archétypes de comportement construits, les consommations d'énergie et les profils de ménages calculés pour chacun des archétypes, on peut produire un descriptif plus fin des styles de vie résidentiels identifiés. Un inventaire des profils calculés est donné dans le tableau 14.

Archétype COV_1 : Maison pratique. Ce type de comportement est associé à des pratiques à forte consommation d'énergie telles que des équipements nombreux pour l'alimentation, l'hygiène et les loisirs, mais aussi à une forte demande de chauffage de chauffage. Par conséquent, ces ménages consomment une quantité importante d'énergie par rapport à l'échantillon total (la consommation médiane est supérieure de 21 % à celle de l'échantillon). Une grande partie des ménages sont des familles très aisées avec enfants, propriétaire d'une maison individuelle assez ancienne située en zone urbaine.

On peut faire l'hypothèse que ces ménages sont plutôt absents du fait de leur activité professionnelle, et qu'ils sont très bien équipés, d'abord du fait de la présence d'enfants, mais aussi du fait de l'absence de contraintes financières mais aussi peut-être d'une culture de consommation importante.

Archétype COV_2 : Maison économe. La plupart des comportements sont caractérisés par la régulation et la modération, y compris le niveau d'équipement et la demande de chauffage. La consommation d'énergie finale (FEC) médiane est donc inférieure à la médiane de l'échantillon. Le profil le plus courant est celui d'un propriétaire occupant célibataire à faibles revenus dans une zone urbaine. Les deux autres profils sont plutôt des locataires de classe moyenne ou des jeunes propriétaires. On peut faire l'hypothèse que ces ménages adoptent des pratiques d'économie d'énergie pour des raisons économiques ou culturelles.

Archétype COV_3 : Maison cocon. Ce type de comportement se caractérise par une très forte présence à domicile, un très fort taux d'équipement et d'utilisation dans tous les domaines (hygiène, alimentation et loisirs). En conséquence, les ménages associés à ce type de comportement ont une En conséquence, les ménages associés à ce type ont une consommation totale d'énergie très élevée (la consommation médiane est de 55% supérieure à celle de l'échantillon). Deux profils de ménages sont représentés. Le premier est celui de jeunes actifs avec enfants, aux revenus moyens, et également propriétaires d'un grand logement en zone rurale. Le second profil est celui de retraités (couples ou célibataires), propriétaires d'un logement de plus de 100 m2 où ils vivent depuis 40 ans. Ces deux profils sont particulièrement intéressants car ils montrent qu'un même comportement énergivore peut être adopté par des ménages très différents. Dans le premier cas, on peut supposer que les ménages se sont équipés pour répondre aux besoins liés à la présence d'enfants. La forte présence à la maison peut être associée à l'absence d'emploi, au travail à domicile et à la présence d'enfants pour des raisons de santé. Un plus grand nombre d'actions d'économie d'énergie les distingue des familles associées à l'archétype 1. Dans le second on peut supposer que les ménages ont conservé les équipements dont ils avaient besoin avant le départ des enfants. De même, la forte présence à la maison peut s'expliquer par le fait que les personnes sont à la retraite.

Archétype COV_4 : Maison chaude. Les ménages de ce type ont des pratiques assez énergivores : l'équipement est élevé pour les loisirs et l'hygiène, il y a une forte demande de chauffage et une faible régulation. Cependant, ces ménages ont tendance à être peu présents au logement. De ce fait, leur consommation médiane d'énergie est supérieure de 24% à celle de l'échantillon total. Pour cet archétype comportemental, il existe une plus grande diversité dans les structures des ménages et des logements associés, puisque cinq profils types ont été construits. On remarque que cet archétype se distingue des autres par une valeur "très élevée" sur la variable VS4 ("Besoin de chauffage") et des comportements de régulation faibles. On peut supposer que le trait commun de ces ménages est une définition du confort associée à une température de chauffage plus élevée et à une consommation non régulée. Cet archétype est particulièrement intéressant car il est moins lié aux contextes résidentiels et il suggère que les

pratiques liés au confort thermique sont moins liées aux contextes résidentiels que les pratiques d'hygiène, d'alimentation, de loisirs, d'occupation du logement.

Archétype COV_5 : En quête de confort. Ce type de comportement se distingue par la prévalence des comportements de régulation (faible équipement, utilisation de LED) et de recherche de confort (équipement de chauffage supplémentaire, obstruction des bouches d'aération). Les ménages associés à ce type de comportement ont une consommation totale légèrement inférieure (médiane inférieure de 13 %) mais une consommation par mètre carré de surface habitable plus élevée (médiane supérieure de 18%), alors que la consommation par personne est très faible (-35%). Il est très intéressant d'observer qu'un très large éventail de profils de ménages est associé à cet archétype. Une grande partie d'entre eux sont célibataires, locataires et appartiennent pour la plupart à des groupes socioéconomiques moyens ou faibles, tandis qu'une autre grande partie est constituée de familles. Une autre grande proportion est constituée de familles avec enfants. Une statistique notable concernant les ménages de cet archétype est qu'ils sont plus susceptibles de vivre dans des zones urbaines, dans des logements beaucoup plus petits (27 m²/hab. contre 37 m²/hab. en moyenne). Par conséquent, ce type de comportement reflète un comportement très restrictif, peut-être pour des raisons financières ou contextuelles (par exemple, associées à une mauvaise qualité de logement).

Archétype COV_6 : Maison frugale. Ce type comprend les ménages peu équipés mais très présents. La quasi-totalité des ménages de ce type sont presque tous des célibataires ou des couples de retraités ayant des revenus moyens ou faibles et vivant dans un appartement en zone urbaine. On peut raisonnablement supposer que ces ménages, pour des raisons financières ou culturelles, adoptent un comportement économe en énergie.

Archétype COV_7 : Maison minimale. Les ménages appartenant à ce type de comportement partagent un ensemble de pratiques d'économie d'énergie telles que l'utilisation de lampes et d'équipements à faible consommation d'énergie. Ce type est très proche du type 6 à l'exception de la rareté des actions écologiques. Quatre profils types sont associés à ce type et décrivent des personnes âgées ou des chômeurs (12 % des ménages de l'archétype contre 5 % dans l'ensemble). Elles vivent principalement seules ou en couple, dans un appartement dont elles sont propriétaires. La frugalité et les contraintes financières pourraient également expliquer cet assemblage de comportements.

Tableau 14 : Inventaire des profils types de ménages identifiés pour chacun des archétypes de comportements. Les profils sont calculés par classification ascendant hiérarchiques des caractéristique des ménages et des logements pour chacun des archétypes de comportement. Les proportions des profils pour chaque archétype sont données en pourcentage. Source : Auteur d'après calculs sur la base ENERGIHAB

Archétype	Profils types des ménages et des logements *La proportion de ménages représentés par chaque profil est donnée en %.*
1	**42%** - Couple à revenu élevé, âgé de 40 à 49 ans, avec enfants. Ils sont propriétaires d'une maison individuelle construite entre 1949 et 1975 dans une zone rurale.
	27% - Couple de cinquantenaires aux revenus moyens avec enfants. Ils sont propriétaires d'un appartement récent de 50 à 75 m² en zone urbaine.
	15 % - Personne âgée célibataire, retraitée, avec des revenus moyens, propriétaire d'un appartement récent de 50 à 75 m² en zone urbaine.
	16% - Couple de retraités, sans enfant à la maison, vivant dans une petite maison ancienne construite avant 1949 dans une zone urbaine.
2	**43%** - Personne active célibataire âgée d'une cinquantaine d'années disposant d'un revenu moyen ou faible. Propriétaire d'un appartement de plus de 100 m² en zone urbaine, construit avant 1949.
	36% - Couple quinquagénaire à revenus moyens, sans enfant, locataire d'un appartement datant d'avant 1949 dans une zone urbaine.
	21% - Jeune famille à revenu moyen avec enfants, propriétaire d'une petite maison récente dans une zone suburbaine.
3	**48%** - Jeune couple avec enfants disposant d'un revenu moyen, propriétaire d'une maison construite avant 1975 en zone rurale, d'une surface habitable comprise entre 75 et 100 m².
	42% - Couple retraité sans enfant avec des revenus moyens ou élevés, propriétaire d'une maison construite avant 1975 dans une zone urbaine, avec une surface habitable comprise entre 75 et 100 m².
	10% - Retraité célibataire à faible revenu, propriétaire d'un appartement construit avant 1949 en zone urbaine, d'une surface habitable comprise entre 75 et 100 m².
4	**23%** - Couple de quinquagénaires avec enfants, locataire d'un logement collectif construit avant 1975 en zone urbaine
	22% - Couple dans la cinquantaine sans enfant, propriétaire d'une maison récente de plus de 150 m² en zone rurale.
	19% - Célibataire d'une cinquantaine d'années, locataire d'une maison construite après 1975 d'une superficie de près de 75 m² dans une zone rurale.
	18% - Couple sans enfant à très hauts revenus vivant dans un logement collectif construit après 1975 dans une zone urbaine
	14% - Couple de retraités, sans enfants, vivant dans une petite maison individuelle en zone urbaine.
	7% - Personne seule avec un revenu moyen dans la trentaine, louant un appartement de moins de 50 m² dans une zone urbaine
5	**26%** - Personne seule, retraitée, avec un très faible revenu et propriétaire d'un appartement de 50 m² dans lequel elle vit depuis 40 ans.
	26% - Jeune couple avec enfants, louant un appartement de 50-75 m² en banlieue
	22% - Couple à hauts revenus avec enfants, propriétaire depuis moins de 5 ans d'un appartement de 75 m² construit entre 1949 et 1975 dans une zone rurale.
	19% - Famille monoparentale avec une personne de référence âgée de 30 à 40 ans, à très faible revenu. Locataire d'un logement dans le parc public.
	7% - Famille à faible revenu avec enfants, où la personne de référence a une quarantaine d'années et est au chômage. Locataire d'une maison individuelle en zone rurale.
6	**65%** - Retraité célibataire de plus de 70 ans, locataire d'un logement collectif construit avant 1949 en zone urbaine
	30% - Couple de retraités à revenu moyen propriétaire d'un appartement récent dans une zone urbaine
	5% - Famille jeune à revenus moyens possédant une maison individuelle construite entre 1949 et 1975 dans une zone urbaine.
7	**43%** - Retraité célibataire, propriétaire occupant d'un appartement de 75 m² en zone urbaine depuis plus de 40 ans.
	27% - Personne seule, âgée de plus de 60 ans et active. Locataire d'un logement collectif en zone urbaine.
	21% - Couple de retraités sans enfants aux revenus moyens ou faibles, vivant dans une maison individuelle en zone rurale.
	9% - Couple de retraités avec un ou plusieurs enfants au foyer. Propriétaire d'une maison individuelle en zone urbaine depuis plus de 30 ans.

Cette analyse qualitative des archétypes permet de dresser un ensemble de situations d'habitation typiques, les styles de vies résidentiels et les consommations d'énergie associées. Là aussi, on constate l'émergence d'archétypes de comportements énergivores, associés à des ménages aisés avec ou sans enfants (archétypes COV_1 et COV_3). A rebours, des ménages composés de personnes seules ou en couple mais enfants et souvent plus âgées adoptent des comportements moins énergivores (COV_2 et

COV_6). La Figure 32 permet d'observer la distribution de chacun de ces archétypes par classe d'âge de la PR et en différenciant les ménages ayant un revenu supérieur ou inférieur à la médiane de l'échantillon.

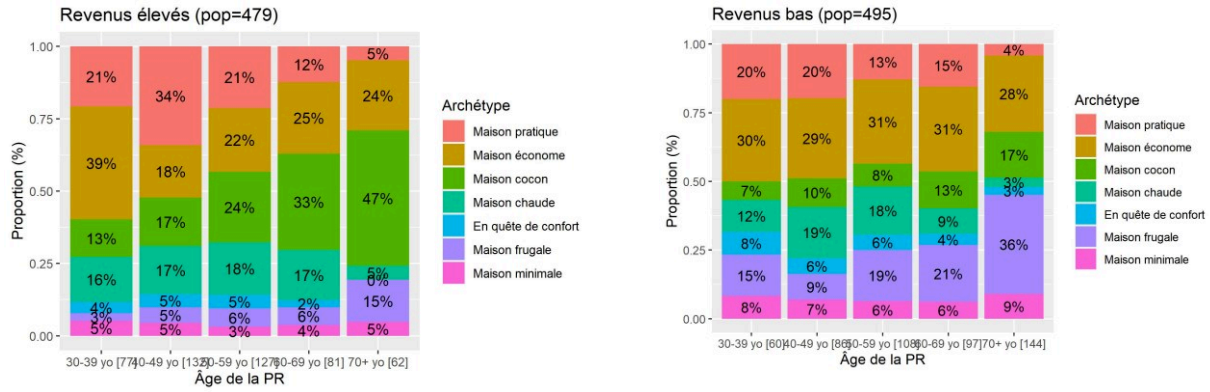
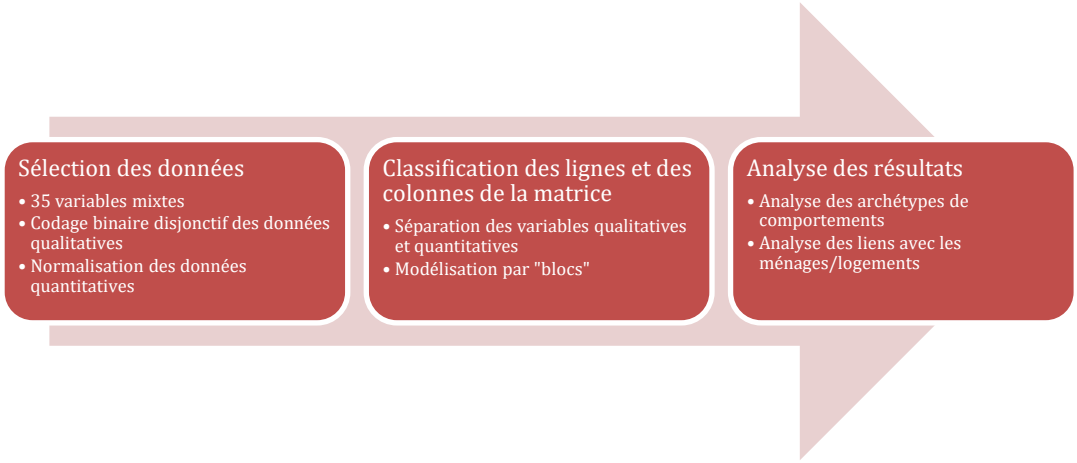


Figure 32 : Evolution des proportions des archétypes de comportements par tranche d'âge. Les archétypes sont calculés à l'aide de la méthode de classification des variables (Données : ENERGIHAB)

On observe deux choses : premièrement on confirme une dépendance entre l'âge de la PR, le niveau de revenu et l'archétype. Aussi, on confirme que les archétypes COV_4 et COV_5 sont peu liés aux caractéristiques des ménages et des logements sélectionnés. Nous reviendrons sur cette remarque dans l'analyse transversale des typologies.

La construction de variables synthétiques de comportement par la méthode ClustOfVar a montré des résultats intéressants notamment en construisant un nombre plus important de variables synthétiques de comportements que la méthode S2 basée sur l'analyse factorielle. Par ailleurs elle a apporté une facilitié dans l'interprétation des variables synthétiques. Dans ce sens, la stratégie S3 apporte selon nous une méthode satisfaisante en alliant interprétabilité, simplicité de calcul et efficacité. Nous souhaitions cependant dans ce travail également comparer l'apport d'une méthode probabiliste pour ce type de travail. L'avantage principal de ce type de méthode est qu'elle ne repose pas sur un critère de distance ou de corrélation pour construire des regroupements des lignes et des colonnes de la base de données mais plutôt sur des hypothèses sur la forme des distributions des variables. Nous décrivons la méthode utilisée et les résultats dans la partie suivante.

2.5 Stratégie S4 : construction d'une typologie par co-clustering de données mixtes



Dans cette partie, nous présentons l'utilisation d'un 4e algorithme, nommé « MixedClust » dans la suite de ce document et présenté dans (Selosse, Jacques et Biernacki 2020). Le « co-clustering » regroupe une famille d'algorithmes de classification non supervisée[24] dont l'objectif est de découvrir des structures latentes associant à la fois les lignes et les colonnes de la table de données. On parle de « Latent Block Modeling ». Dans cette partie, notre intérêt est de changer de critère de classification par rapport aux stratégies précédentes. Celui-ci est non plus géométrique mais probabiliste : ce n'est pas la distance des individus au centre de la classe qui compte mais les distributions des données. L'hypothèse derrière cette approche est que les données traitées sont issues de K blocs de données, chacun caractérisé par une distribution paramétrée (loi normale, lois de Bernoulli etc.) et que chaque individu peut être rattaché à un mode par une probabilité d'appartenance. On peut considérer cette approche comme une extension des modèles usuels de « mélange » (en anglais « mixture models ») qui propose une approche de classification non supervisée où les appartenances ne sont pas strictes mais définies par des degrés d'appartenance. Les modèles de mélange ont d'abord été développé pour réaliser des classifications par ligne des tables de données. Nous nous intéressons ici à un algorithme qui effectue une classification simultanée des lignes et des colonnes.

L'intérêt de cette méthode est double ici : d'abord elle permet regrouper les deux étapes de calcul (classification des colonnes et classification des lignes) et donc de limiter les biais de calcul introduits par le modélisateur[25]. Ensuite, cette méthode utilise un critère probabiliste pour construire des blocs homogènes de lignes et de colonnes. Ce critère probabiliste diffère ontologiquement des méthodes géométriques car il suppose dans sa formulation que les réponses des répondants et par extension les

[24] Le terme « non supervisé » utilisé dans le champ des sciences de la donnée désigne le fait d'entraîner un modèle, c'est-à-dire calculer la valeur de ses coefficients, sans pour autant connaître l'état de la variable expliquée. A l'opposé, l'apprentissage dit « supervisé désigne le fait de calculer ces coefficients en fournissant à l'algorithme des exemples de cas où sont connus les états des variables explicatives (dépendantes) et de la variable expliquée (indépendante). L'entrainement en apprentissage « non supervisé » repose donc sur des critères quantitatifs à définir par le modélisateur.

[25] En effet, si on prend l'exemple de la méthode S2 associant une analyse factorielle avec une CAH, on observe que le choix du nombre de facteurs principaux retenus par le modélisateur influence directement le calcul de la CAH.

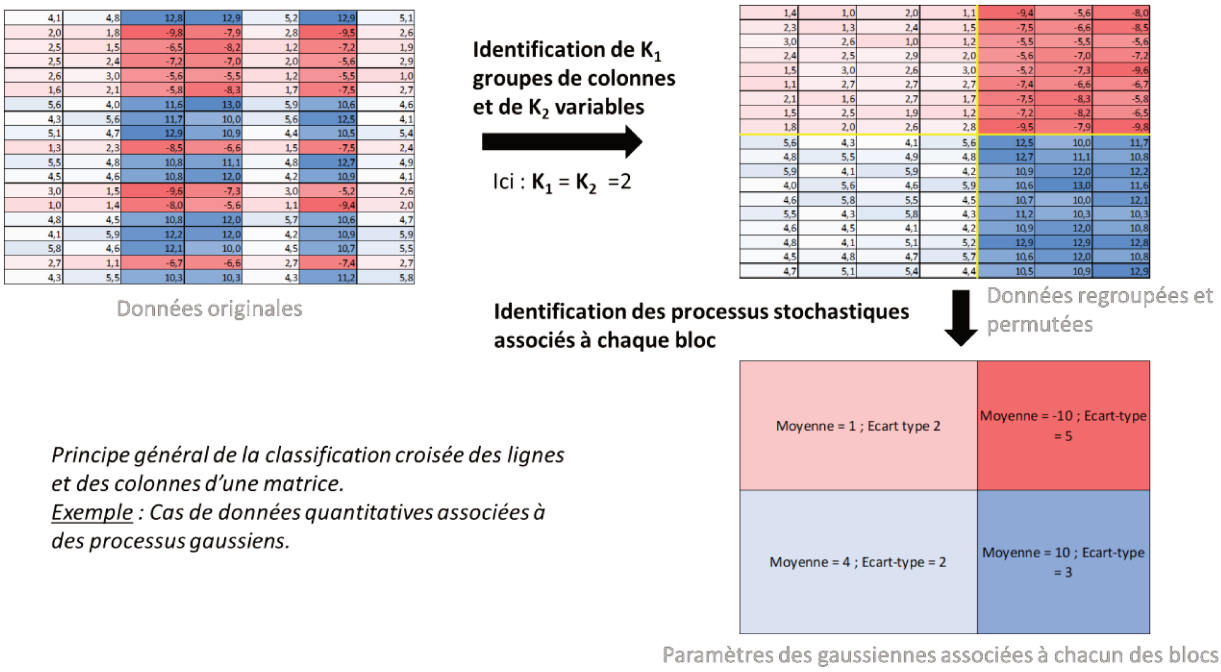
valeurs prises par les variables des réalisations de distributions de probabilités. La formulation de l'algorithme de co-clustering vise ainsi à restituer cette perspective. Concrètement, l'algorithme suppose que les variables qui ont des valeurs à des niveaux comparables de manière récurrentes ont une probabilité élevée d'être issue d'une loi de probabilité commune. Par exemple, si on suppose que le nombre d'équipement pour l'hygiène et pour l'alimentation sont deux variables qui renvoient à un même processus stochastique gaussien, les deux variables seront regroupées dans un même groupe de variables et partageront les paramètres d'une même loi gaussienne. Dans la première partie, nous décrivons brièvement la méthode « MixedClust » à l'aide d'un exemple simple puis nous présentons les résultats de son application à nos données.

Méthodologie

Présentation de l'algorithme de co-clustering MixedClust de M. Selosse

Un exemple introductif

Pour faciliter la compréhension du lecteur, nous proposons ici un petit exemple introductif (Figure 33). Nous avons généré des données à partir de lois de probabilité (gaussiennes), selon une structure de blocs : les 9 lignes forment un premier groupe et les 10 autres un second. De même, nous avons créé deux groupes de colonnes, un de 4 et l'autre de 3 colonnes. Ce faisant, 4 blocs sont créés et des données sont générées à l'aide de lois de probabilités gaussiennes. Nous avons ensuite mélangé les lignes et les colonnes pour obtenir l'ensemble de « données originales » sur la figure. L'enjeu d'un algorithme de co-clustering est alors de parvenir à retrouver 3 choses : les groupes de lignes, les groupes de colonnes et les paramètres des lois gaussiennes associées à chacun de ces blocs. Il est à noter que dans cet exemple, les données sont uniquement quantitatives. Dans notre travail, l'utilisation de données mixtes impose de recourir à un algorithme légèrement plus complexe qui réalise cette même classification mais en séparant les données quantitatives d'un côté (qui sont supposées obéir à des lois gaussiennes) et qualitatives de l'autre (qui sont supposées être régies par des lois multinomiales).



Principe général de la classification croisée des lignes et des colonnes d'une matrice.
Exemple : Cas de données quantitatives associées à des processus gaussiens.

Figure 33 : Exemple introductif pour comprendre le principe du co-clustering. Source : Auteur

La formulation mathématique du problème de classification croisée

Le modèle de « MixedClust » a pour fonction de calculer simultanément : une partition des lignes, une partition des colonnes et les paramètres associés aux lois de probabilités de chacun des blocs. Le critère qui guide la recherche des différents éléments est la maximisation de la vraisemblance du modèle. Celle-ci est définie comme la probabilité d'observer les données de l'enquête, conditionnellement à un paramétrage :

$$L(\mathcal{M}) = p(\boldsymbol{X}\,|\boldsymbol{\theta}) = \prod_{i,j} p\left(x_{i,j}\,|\boldsymbol{\theta}\right)$$

Où i, j désignent les indices des lignes et des colonnes dans la matrice de données $\boldsymbol{X}$. θ désigne les paramètres du modèle $\mathcal{M}$. La densité de probabilité p dépend de la nature de la variable $x_{.j}$: elle désigne soit une loi normale avec deux paramètres (moyenne, écart type) dans le cas d'une variable quantitative, sinon une loi multinomiale de paramètre m désignant le nombre de modalités, et les probabilités p_m de chacune d'entre elles.

En pratique, ce type de modèle est entrainé en maximisant le logarithme de la vraisemblance. Dans leur article de 2020 (Selosse et *al.*, 2020), les auteurs rappellent l'intérêt d'utiliser une version stochastique de l'algorithme EM et introduisent le critère ICL (acronyme de l'anglais « Integrated Complete-data Likelihood ») qui est très utilisé dans la littérature pour entraîner de tels modèles (Biernacki, Celeux,et Govaert 2000). Nous ne développons pas dans ce paragraphe les détails techniques liés à l'entraînement de l'algorithme car leur complexité nécessiterait un développement relativement long. Les lecteurs

trouveront toutefois les détails techniques dans l'article de référence (Selosse, 2020). Pour le calcul, nous utilisons la librairie *mixedClust* sur R[26]. Le paramétrage du modèle repose sur 5 paramètres :

- K_r : le nombre de clusters de lignes. Il s'agit là du nombre d'archétypes de comportement.
- K_{c1} : le nombre de clusters colonnes pour les variables quantitatives uniquement.
- K_{c2} : le nombre de clusters colonnes pour les variables qualitatives uniquement.
- m : le nombre de modalités. Il est utilisé pour le calcul de chacune des variables aléatoires multinomiales synthétiques.

Le choix de ces paramètres ne peut pas être fait de manière itérative car le choix d'un paramètre influence nécessairement le choix optimal (au sens du critère ICL) des autres. La recherche du paramétrage optimal a été faite par « recherche brute » c'est-à-dire testant toutes les combinaisons possibles des variables pour les intervalles suivants des variables :

- $K_r \in [\![1, 14]\!]$
- $K_{c1} \in [\![2, 9]\!]$
- $K_{c2} \in [\![2, 9]\!]$
- $m \in [\![2, 7]\!]$

Le calcul de chacun de ces 5376 combinaisons de paramètres a été fait uniquement une fois en raison du coût computationnel important. Ce choix fragilise nos résultats car l'entrainement du modèle ne converge pas nécessairement vers le modèle optimal (dépendance à l'initialisation aléatoire du modèle). Toutefois, le grand nombre de modèles calculés nous permet d'identifier une série de modèles performant. Le modèle retenu est le modèle qui maximise le critère ICL.

Analyse des résultats

La recherche du modèle maximisant le critère ICL a fourni une série de modèles ayant un ICL maximal. Nous donnons ici à titre indicatif les 6 modèles (Tableau 15) et les paramétrages sélectionnés ayant les critère ICL les plus élevés à l'issue de la phase de recherche du modèle optimal.

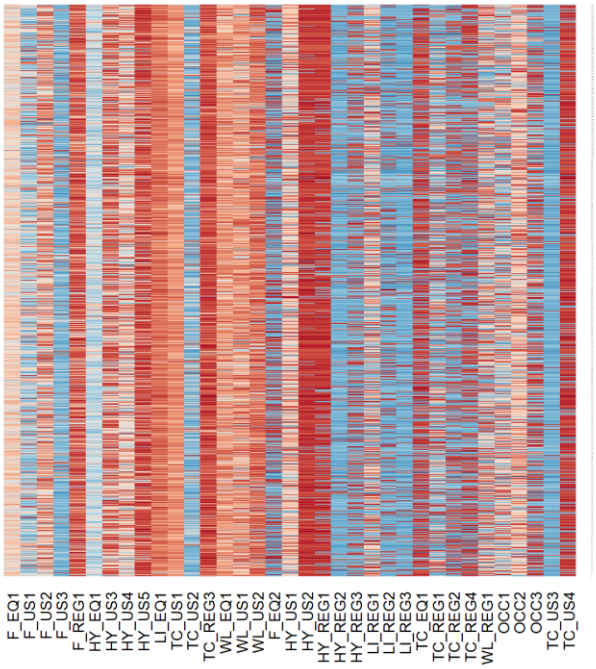
Tableau 15 : Liste des 6 premiers modèles présentant les meilleures performances (au sens de l'indice ICL).

Modèle	ICL	K_r	K_{c1}	K_{c2}	m
1	-40274	12	7	8	7
2	-40856	14	7	9	6
3	-41082	7	6	8	4
4	-41103	11	5	6	6
5	-41164	5	9	8	5
6	-41201	11	6	9	3

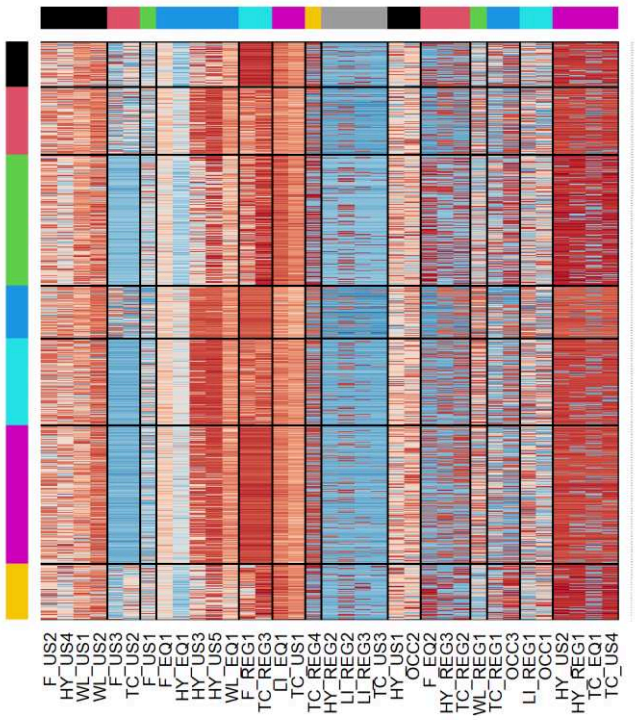
[26] La librairie *mixedClust* contient les fonctions nécessaire à l'entraînement et l'analyse du modèle de co-clustering de données mixtes de (Selosse, 2020). Elle est disponible au lien suivant : https://cran.r-project.org/web/packages/mixedClust/index.html

On observe que parmi ces « meilleurs » modèles, on observe que tous on tendances à construire au moins 5 groupes de colonnes quantitatives (K_{c1}) et 6 groupes de colonnes qualitatives (K_{c2}). En revanche on observe deux tendances différentes : 4 modèles proposent de distinguer un nombre important de clusters de lignes (K_r supérieur ou égale à 11) tandis que 2 autres proposent de ne distinguer que 5 ou 7 archétypes.

Pour faciliter la présentation des archétypes et l'appréhension de la méthode nous proposons ici de retenir le modèle 3 qui présente 7 archétypes de comportement. Une représentation de la matrice de données initiale est donnée à la Figure 34 (A). Chaque ligne comprend les valeurs prises par les variables de comportement pour un ménage. Chaque colonne est composée de l'ensemble des valeurs prises par une variable de comportement. Le symbole des variables est donné pour chaque colonne mais le numéro des lignes a été supprimé pour préserver l'anonymat. Un code couleur est choisi pour pouvoir représenter les variations relatives de chacune des variables. La même table de donnée est à nouveau représentée Figure 34 (B), cette fois en réorganisant les colonnes et les lignes de manière à former les groupes de lignes et de colonnes identifiés par la modèle retenu. Dans cette seconde figure, il est possible d'identifier une homogénéité plus importante des couleurs par bloc, témoignant d'une proximité à la fois entre les colonnes et les variables. Une analyse plus approfondie des regroupements entre les colonnes et les lignes est effectuée afin de mieux comprendre quelles sont les « dimensions comportementales » repérées par cette approche (en observant les regroupements entre variables) et quels sont les archétypes de comportement calculés par cette méthode ?



(A)



(B)

Figure 34: Tracé des tableaux de données ENERGIHAB. Les données originales, non classées sont données sur la figure (A). Les données réorganisées en lignes et en colonnes sont données sur la figure (B). Les données sont mises à l'échelle entre 0 et 1 pour permettre une représentation graphique exploitable. Seule l'homogénéité des couleurs au sein des bloc est intéressante. Source : Auteur, calculs effectués sur la base ENERGIHAB.

L'analyse des regroupements des variables permet de voir qu'il y a 14 groupes de variables, ce qui est plus important que dans les stratégies S2 et S3. Il faut toutefois garder à l'esprit que cette méthode n'a pas permis de regrouper les variables qualitatives et quantitatives et il est probable qu'une même thématique (par exemple le niveau d'équipement) puisse être caractérisé par un groupe de variables qualitatives et un groupe de variables quantitatives. On peut étudier les ensembles de variables pour caractériser les dimensions saillantes. On étudie les groupes de gauche à droite sur la Figure 34B. On parle ici de « groupe de variables » plutôt que de variable synthétique car chaque groupe n'est pas caractérisé par une variable quantitative. L'interprétation donné au groupe de variable est faite à partir du croisement des variables composant le groupe. Pour rappel, nous n'utilisons ici que les symboles des variables (par F_EQ1) et leur description peut être trouvée à la page 82. On recense les groupes de variables dans le tableau 16.

Tableau 16 : Liste des groupes de variables identifiés par la stratégie S4. Source : Auteur après calculs sur la base PHEBUS.

Nom du groupe de variables	Variables initiales
Groupe 1 (Quantitatif) : Usage des appareils domestiques	F_US2; HY_US4; WL_US1; WL_US2
Groupe 2 (Quantitatif) : Présence au domicile et ventilation	F_US3; TC_US2
Groupe 3 (Quantitatif) : Prise de repas au domicile	F_US1
Groupe 4 (Quantitatif) : Niveau d'équipement du ménage	F_EQ1; HY_EQ1; HY_US3; HY_US5; WL_EQ1
Groupe 5 (Quantitatif) : Gestes verts et comportements d'économie	F_REG1; TC_REG3
Groupe 6 (Quantitatif) : Demande en chauffage	LI_EQ1; TC_US1
Groupe 7 (Qualitatif) : Comportements d'économie du chauffage	TC_REG4
Groupe 8 (Qualitatif) : Gestes verts et comportements d'économie	HY_REG2; LI_REG2; LI_REG3; TC_US3
Groupe 9 (Qualitatif) : Présence au domicile	HY_US1; OCC2
Groupe 10 (Qualitatif) : Gestes verts	F_EQ2; HY_REG3; TC_REG2
Groupe 11 (Qualitatif) : Régulation des consommations numériques	WL_REG1
Groupe 12 (Qualitatif) : Présence moyenne le WE	TC_REG1; OCC3
Groupe 13 (Qualitatif) : Intensité d'usage du logement pour les activités domestiques	LI_REG1; OCC1
Groupe 14 (Qualitatif) : Demande en ECS et en chaleur	HY_US2; HY_REG1; TC_EQ1; TC_US4

L'analyse des groupes de variables permet de faire plusieurs observations. En premier lieu, on remarque que les groupes sont plus nombreux avec cette approche que dans les autres stratégies. Trois groupes ne sont composés que d'une seule variable et plusieurs groupes semblent communiquer des informations sur les mêmes dimensions comportementales (présence au logement, équipement, régulation). Selon notre expérience, il semblerait que la méthode présente des limites techniques pour regrouper les variables. Nous avons fait le choix à ce stade de ne pas creuser ce point mais il nous semble que c'est là un point d'amélioration crucial de la méthode, pour lequel des propositions de modélisation pourraient être faites.

Le second temps de l'analyse vise à étudier les archétypes de comportements, et les logements et ménages associés. De manière similaire aux paragraphes précédents, on propose de lier ces descriptions pour faciliter l'interprétation et la compréhension des résultats de classification.

- **Archétype COCL_1 « Maison économique » :** ce cluster comprend des ménages qui sont des couples d'actifs, plutôt sans enfant, aux revenus moyens et propriétaires de leur logement. En termes de comportements domestiques, on observe qu'en moyenne ces ménages ont un équipement alimentaire moyen mais des équipements et des usages de loisirs numériques importants. Actifs, ils ont une présence moyenne au domicile, mais ont des comportements de régulation du chauffage importants. Ces ménages ont une consommation en énergie finale totale moyenne par rapport à l'échantillon.
- **Archétype COCL_2 « Maison frugale » :** ce groupe comprend des ménages qui sont composés plutôt de personnes seules, de tout âge et ayant des revenus moyens à faibles. Ils occupent très souvent un logement collectif dont ils sont locataires. En termes de comportements, on observe que

ces ménages ont des niveaux d'équipement très faibles excepté en termes de loisirs numériques, une occupation du logement importante, et des gestes de régulation du chauffage très importants.

- **Archétype COCL_3 « Maison confortable » :** cet archétype est caractérisé par un équipement très élevé et des gestes de régulation des consommations très bas. Les ménages rattachés à ces pratiques sont essentiellement des couples avec enfants aux revenus moyens voire élevés et vivant en tant que propriétaire occupant de leur logement souvent individuel. Les consommations en énergie finale totale sont en moyenne très élevées en comparaison de l'échantillon.
- **Archétype COCL_4 « Cocon frugal »** : cet archétype est caractérisé par les taux d'équipements et d'usage les plus faibles entre tous, ainsi qu'une occupation moyenne du logement plus importante. Les ménages rattachés à cet archétype sont à 70% des ménages composé d'une personne seule aux revenus situés dans le premier quintile et vivant comme locataire, souvent dans le parc public, dans un logement collectif de moins de 70 m². Une autre caractéristique est que les logements occupés par ces ménages sont relativement à l'échantillon total, deux fois moins souvent rénovés. In fine, ces ménages ont des consommations d'énergie finale totale très faibles.
- **Archétype COCL_5 « Maison froide »** : cet archétype de comportement est caractérisé principalement par un bas niveau d'équipement et des pratiques de chauffage très basses. Un point intéressant est que cet archétype renvoie à des situations d'habitation plutôt diverses. Un point commun entre elles est leur revenu moyen ou faible et le fait que les logements occupés sont souvent des appartements de moins de 70 m².
- **Archétype COCL_6 « Cocon chaud »** : cet archétype se caractérise par une forte présence au domicile et la faiblesse des gestes de régulation thermique. Il s'agit dans ce cas de ménages souvent plus âgés, aux revenus faibles ou moyens et occupant leur logement en tant que propriétaires.
- **Archétype COCL_7 « Maison des loisirs** : cet archétype est caractérisé par un très haut niveau d'équipement dans tous les domaines de consommation, mais des gestes de régulation aussi très importants. Peu présent dans leurs logements, les ménages associés sont en fait essentiellement des ménages composé de deux actifs avec 2 ou 3 enfants, avec des revenus moyens ou élevés et très souvent propriétaire d'une maison individuelle. Les consommations en énergie finale totale de ces ménages sont très élevées.

L'analyse des distributions des consommations d'énergie pour chacun des indicateurs montre à nouveau que la classification des comportements à partir de la Figure 35 permet de différencier les consommations en énergie finale totale mais pas les distributions par m² et par personne. Ce constat appuie l'idée que l'étude de l'association des comportements domestiques renvoie d'abord à des contextes résidentiels.

La réalisation de ce 4^{e} calcul montre des rapprochements avec les précédents. Il est intéressant à présent de croiser les résultats de chacune des approches. Ce croisement permet d'étudier l'intérêt relatif des stratégies et de produire des résultats plus robustes.

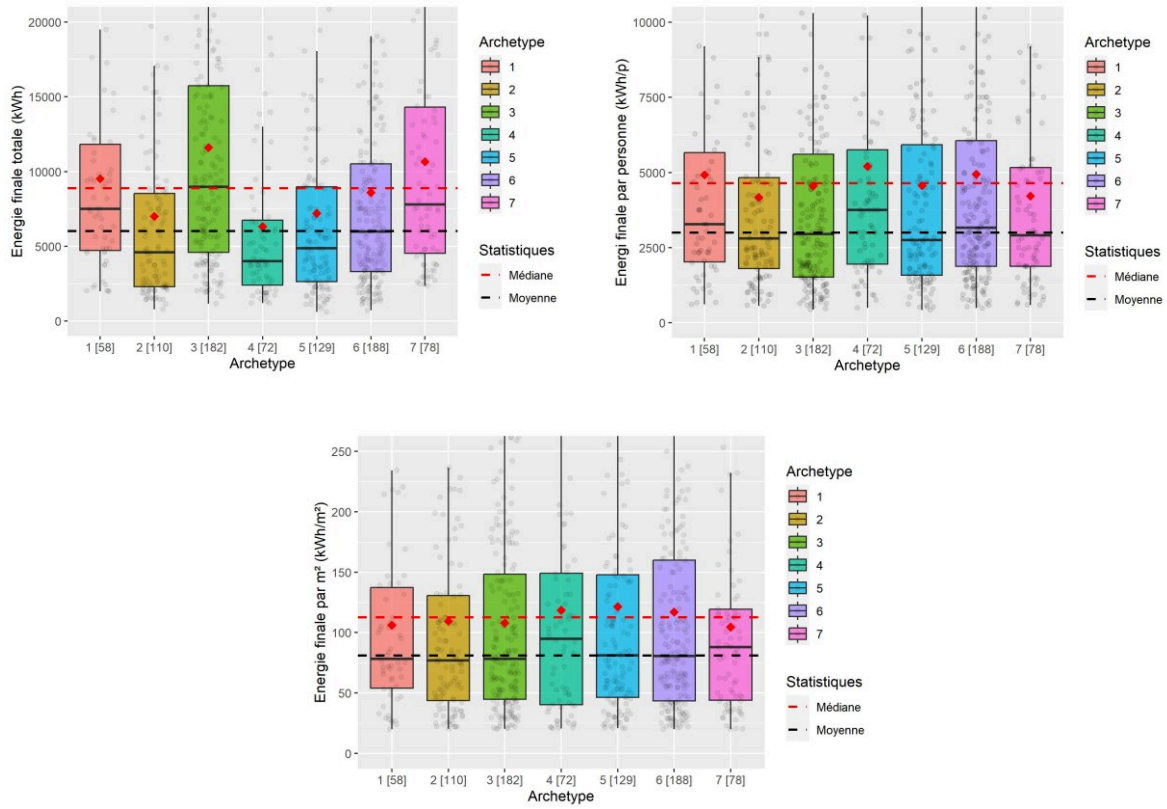


Figure 35 : Distribution des consommations en énergie finale totale (A), par personne (B) et par mètre carré (C) pour chacun des archétypes de comportement. Source : Auteur après calculs sur la base ENERGIHAB.

3. Comparaison des approches de classification

La comparaison des approches de classification permet de mieux comprendre à la fois les résultats de classification obtenus. Nous comparons dans cette partie les classifications des ménages selon leurs pratiques domestiques. Nous présentons un diagramme de Sankey (Figure 36) qui permet de comparer les quatres stratégies de classification (S1, S2, S3, S4).

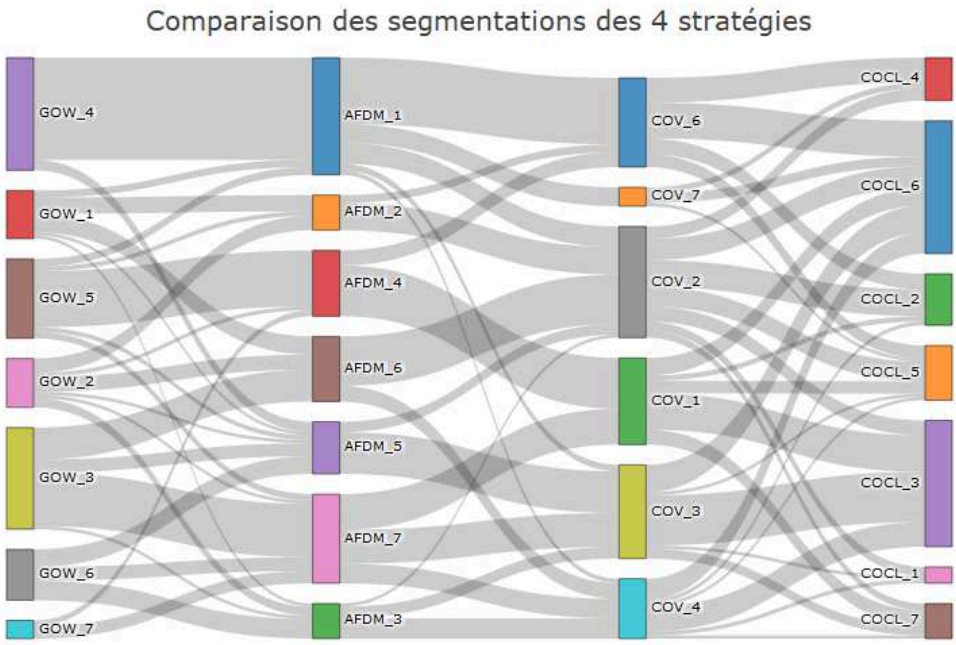


Figure 36 : Diagramme de Sankey pour les 4 classifications calculées dans ce chapitre. Le diagramme permet de visualiser les recouvrements et les divergences en termes de classification des 4 stratégies. Les flux en gris ont une largeur proportionnelle au nombre de ménage. Un flux épais entre deux classes en couleur (par exemple GOW_4 et AFDM_1) témoigne du fait que les algorithmes des stratégies S1 et S2 ont tous deux regroupé un grand nombre de lignes de la base de données au sein d'un même segment. Source : Auteur.

La comparaison des classifications des individus permet de faire plusieurs remarques. En premier lieu, les classifications n'aboutissent pas à des segmentations identiques puisque les « branches » reliant les segments (en couleur sur la figure) décrivent plutôt une arborescence que des liaisons simples. En revanche, certaines branches sont plus épaisses sur ce graphique et témoignent du fait que certains individus se retrouvent au sein d'un même cluster pour plusieurs stratégies de classification. A titre d'exemple, on voit que la branche grise reliant le cluster « GOW_4 » et le cluster « AFDM_1 » est très épaisse et il n'existe que trois branches résiduelles liant « GOW _4 » à « AFDM_5 », « AFDM_1 » à « « GOW_1 » et « AFDM_1 » à GOW_5 ». Cette observation permet de dire que les clusters « GOW_4 » et « AFDM_1 » sont presque identiques alors qu'ils sont issus de deux calculs différents. Dans une approche probabiliste, on peut dire que ce résultat augmente la significativité du cluster construit. Sur le plan méthodologique, il invite à dire que les deux méthodes de classification sont proches (puisqu'elles fournissent des résultats similaires) mais pour conclure il faut examiner de plus près les autres liens entre les clusters.

En remarque liminaire, on peut dire que la stratégie S4 offre une classification très différente des autres stratégies. L'indice de Rand ajusté (ARI pour *Adjusted Rand Index*) permet de quantifier le degré de similarité entre deux classifications. Il est défini comme la proportion de liens communs entre les éléments pour deux classifications données, corrigé de l'espérance. Ainsi l'indice ARI vaut 0 pour deux clusterings indépendants et 100% pour deux clusterings identiques. On renseigne dans la Figure 37 l'indice de Rand ajusté entre les différents clusterings calculés. L'analyse des coefficients montre que

les classifications ne sont pas globalement similaires car l'indice ARI varie entre 4,3 et 18,3%. On remarque aussi que le clustering issu de la stratégie S4 diffère significativement des clusterings issus de S1 (ARI=6,7%°) et de S3 (ARI=4,3%). Par souci de simplicité, et au vu de sa singularité par rapport aux autres méthodes nous l'écartons pour l'instant de la suite des analyses. La stratégie S2 semble jouer un rôle central parmi les différentes stratégies puisque c'est elle qui partage le plus grand nombre de similarités avec les autres stratégies ($ARI_{S2,S1} = 18{,}3\%$, $ARI_{S2,S3} = 17{,}1\%$, $ARI_{S2,S4} = 11{,}2\%$).

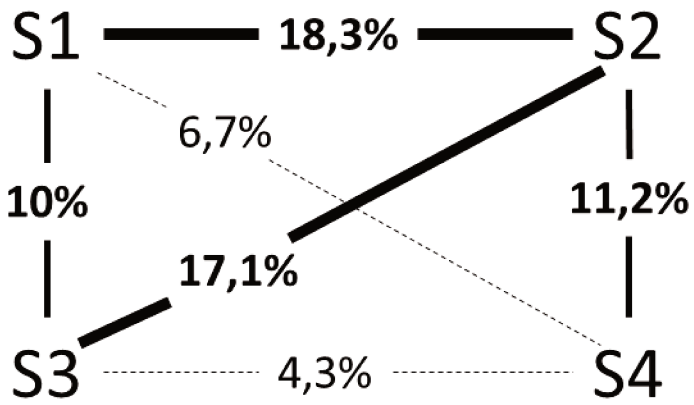


Figure 37 : Indices de Rand Ajustés entre les 4 classifications. Les traits liants les 4 stratégies (S1, S2, S3, S4) ont une épaisseur proportionnelle à l'indice ARI. Source : Auteur après calculs sur la base ENERGIHAB.

Cette comparaison globale permet de mettre en évidence l'intérêt de l'analyse factorielle de données mixtes comme une méthode robuste par rapport aux autres approches explorées ici. Ainsi, si son analyse est plus difficile et que les axes sont plus parcimonieux elle se révèle ici être une méthode de référence intéressante.

En observant le digramme de Sankey, il paraît cependant difficile de tirer un message simple alors les flux semblent imbriqués. Nous avons proposé de réaliser une synthèse des styles de vie résidentiels identifiés en regroupant les classes en quatre styles de vies, témoignant du recoupement des résultats des quatre algorithmes. Ce regroupement est effectué de manière qualitative en observant les liaisons entre les différentes classes. Une synthèse graphique est donnée à la Figure 38. Les quatre styles de vie résidentiels se différencient en termes de comportements en termes de présence, de pratiques de régulation et d'équipement. Une présentation synthétique des quatre styles de vie est proposée.

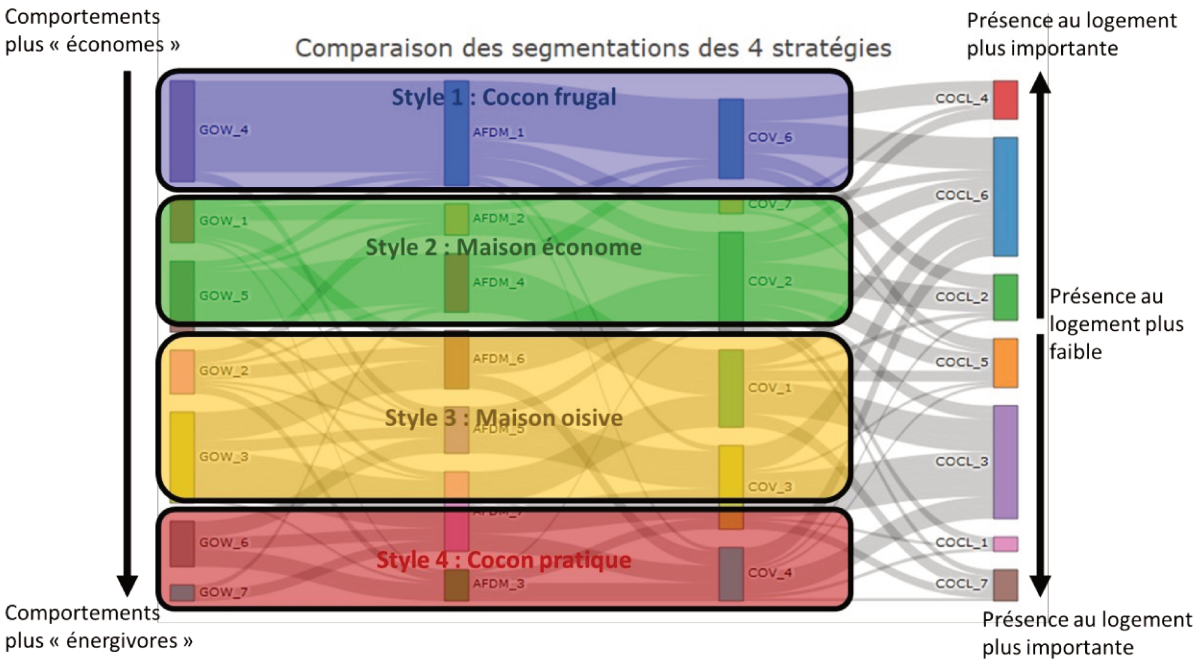


Figure 38 : Identification de 4 styles de vie résidentiels à partir du recoupement des 3 segmentations des comportements issus des stratégies S1, S2 et S3. L'identification est faite de manière qualitative à partir de l'observation des liaisons principales entre les classes. Source : Auteur après calculs sur la base ENERGIHAB.

- **Un premier style de vie résidentiel : le « Cocon frugal »**

L'observation transversale des classifications permet d'observer que les classes GOW_4 ; AFDM_1, COV_6 et COV_7 regroupent les mêmes éléments de la base de données. L'observation des comportements domestiques dans chacune de ces classes a montré un nombre important de gestes économes, un faible équipement et une présence moyenne à importante au logement. Les ménages associés à ces quatre groupes sont en fait presqu'exclusivement des ménages retraités, composés de personnes seules ou de couples âgés et sans enfants, issus de classes ouvrières et ayant des revenus faibles. Leurs consommations d'énergie sont inférieures à la moyenne selon les trois indicateurs de consommation retenus.

- **Un second style de vie résidentiel : la « Maison économe »**

Le second style de vie résidentiel que nous identifions est celui de jeunes couples avec ou sans enfants ayant des revenus moyens, vivant comme locataire ou primo-accédant dans un logement plus petit que la moyenne des ménages de taille identique. Ajoutée à la faible taille du logement, les ménages y ont des comportements générant peu de consommations d'énergie : faible présence (due à l'activité professionnelle vraisemblablement), faible équipement et nombreux gestes de régulation. Ces ménages ont des consommations énergétiques globalement faibles. Les principales classes rattachées à ce type sont les classes AFDM_2, AFDM_6, COV_2, GOW_1 et GOW_2.

- **Un troisième style de vie résidentiel : la « Maison oisive »**

Le troisième style de vie résidentiel identifié regroupe les classes GOW_3, GOW_5, AFDM_4, AFDM_7, COV_1 et COV_4. Les comportements adoptés dans par les ménages rattaché ce style de vie sont les suivants : un équipement alimentaire et d'hygiène moyen à important, un équipement de loisirs plutôt important, un présence moyenne, des gestes de régulations plutôt faibles. Les ménages rattachés à cette situation d'habitation sont plutôt de jeunes couples avec enfants, aux revenus moyens à élevés et occupant un logement plus souvent individuel qu'ils possèdent.

- **Un quatrième style de vie résidentiel : le « Cocon pratique »**

L'analyse transversale des typologies permet d'observer un dernier style résidentiel qui associe un ensemble de pratiques énergivores : un fort équipement, une présence au logement importante, peu de gestes de régulation. Ce style recoupe principalement les classes GOW_6 et GOW_7, AFDM_3 et AFDM_5, COV_3. Une remarque intéressante est que les situations d'habitation reliées à ce style de vie sont à la fois des ménages plutôt jeunes et des ménages en fin de cycle de vie, mais qui occupent presque tous une maison individuelle qu'ils possèdent. Ainsi, on peut faire l'hypothèse que les motivations des jeunes ménages avec enfants pour avoir un fort équipement alimentaire sont d'abord liés à des besoins liés à la structure du ménage (présence des enfants, besoin de gagner du temps dans la réalisation des activités quotidiennes). Pour les ménages plus âgés le fort niveau d'équipement peut être lié soit à l'accumulation d'équipement ou à une logique de consommation par exemple. On remarque cependant que l'essentiel des ménages rattachés à ces classes ont des revenus élevés voire très élevés.

Ce redécoupage en quatre styles de vie résidentiel n'est pas parfait : on peut observer le lien fort entre les classes AFDM_4 et COV_1 et AFDM_6 et COV_2 qui témoigne bien d'une perméabilité entre les quatre styles que nous venons de donner. Cette définition permet toutefois de rendre compte d'une manière robuste de tendances globales en termes d'occupation, d'équipements, de comportements de régulation.

Les groupes construits peuvent être comparés dans une certaine mesure aux classifications de la littérature. Les style de vie résidentiel numéro 3 et 4 peuvent être reliés aux classes "spenders" de Van Raaij et "active spenders" de Ben et Steemers. Toutefois, le fait d'étudier les situations d'habitation nous permet de différencier les ménages retraités avec une forte présence à domicile (style numéro 4) et les ménages actifs avec des enfants et un taux d'occupation du domicile plus faible (style de vie numéro 3). Les "conservatives" des auteurs précédents coïncident avec les styles de vie numéro 1 et 2, en différenciant là aussi des ménages plutôt pauvres et/ou âgés (style 1) et des ménages plus jeunes (style 2).

4. Conclusion du chapitre

L'objectif ce chapitre était de produire une analyse des comportements des ménages en lien avec les caractéristiques des situations d'habitation, en vue de produire un modèle de consommation intégrant les contextes et les pratiques des ménages. Le travail a été décliné en trois temps. Une première partie avait pour objectif de décrire les données de comportements de l'enquête ENERGIHAB. Elle a permis de présenter une première méthode de modélisation des comportements basée sur un critère de corrélation. La comparaison des résultats obtenus avec une analyse factorielle de données mixtes montre que cette dernière présente une meilleure stabilité avec l'échantillon, mais qu'elle est moins précise parce qu'elle produit moins de facteurs identifiables. L'analyse des variables synthétiques a permis d'identifier les dimensions comportementales suivantes : niveau d'équipement alimentaire, présence moyenne au logement, niveau d'équipement et intensité d'usage en matière d'hygiène domestique, niveau d'équipement et intensité d'usage des appareils de loisirs numériques, demande en chauffage, absences longues, niveau de régulation du chauffage, pratiques de restriction.

Dans un second temps, le travail a interrogé la possibilité d'étudier les comportements simultanément ainsi que supposé dans le cadre de modélisation proposé au chapitre 1. Cette étude des styles de vie résidentiels suppose sur le plan mathématique de réaliser des travaux de classification de données mixtes (qualitatives et quantitatives). Un résultat important est qu'un lien statistique fort a été mis en évidence entre les comportements et entre les styles de vie résidentiels et les situations d'habitation. Une discussion sur la sensibilité des résultats à quatre stratégies de classification a été menée et quatre styles de vie résidentiels ont finalement été identifiés en croisant les résultats des quatre stratégies.

- Le style de vie « Cocon frugal » caractérise des ménages aux revenus faibles, locataires, plutôt âgés et retraités, ayant des pratiques de régulation thermique importantes et occupant des logements plus petits que la moyenne.
- Le style de vie « Maison économique » renvoie quant à lui à des ménages en début de cycle de vie et aux revenus moyens. Ceux-ci ont des équipements plus nombreux que le style précédent, et sont moins présents, mais adoptent également des gestes de régulation importants
- Le style de vie « Maison oisive » a été caractérisé comme un style de vie où la présence au logement est moyenne, l'équipement important, l'intensité d'usage élevée, et les gestes de régulation sont peu nombreux. Les ménages rattachés à ce style sont surtout aisés et propriétaires d'une maison individuelle.
- Le style de vie « Cocon pratique » traduit un ensemble de gestes énergivores (peu de gestes de régulation, forte présence au logement, fort équipement et intensité d'usage). Les ménages rattachés à ce style de vie sont surtout en fin de parcours résidentiel (couple sans enfants, retraités), aisés et occupant un grand logement (souvent individuel).

L'identification de ces quatre styles de vie permet de conforter le cadre de modélisation proposé au chapitre 1 : il confirme en effet le fait que la classification de données de comportements, d'équipement permet d'identifier des agrégats interprétables (les styles de vie résidentiels), qui sont également cohérents avec les situations d'habitation. Ce lien n'est toutefois pas déterministe comme le montrent les écarts entre les différentes classifications. Dans ce cadre, des travaux complémentaires doivent être menés afin de les mettre en cohérence. On pourra citer par exemple et de manière non exhaustive :

- Intégrer des variables caractérisant le sens attribué aux pratiques domestiques afin de classifier l'ensemble des variables caractérisant les systèmes énergétiques domestiques (voir chapitre 1 pour la définition).
- Intégrer à l'analyse des caractéristiques individuelles pour mesurer l'influence des déterminants individuels ou de contexte (structure du ménage, parcours résidentiel, quartier) sur le système énergétique domestique.

En l'état, les résultats de ce chapitre nous invitent à considérer une très forte liaison entre les pratiques domestiques et les situations d'habitation. Les données de l'enquête ENERGIHAB n'étant pas suffisamment précises sur les consommations en énergie finale, nous utiliserons dans le chapitre suivant les données issues de l'enquête nationale française PHEBUS (2012). Nous ferons l'hypothèse que nos résultats restent valables et nous décrirons dans le chapitre suivant la méthode pour construire un modèle de consommation articulant situations d'habitation et pratiques énergétiques.

Chapitre 3. Construction d'un modèle de la consommation d'énergie domestique à partir des situations d'habitation

Dans le chapitre précédent, nous avons montré que les variables caractérisant les logements et les ménages étaient très liées aux variables décrivant les comportements domestiques (i.e. modélisant les pratiques sociales). En particulier, nous avons montré qu'il était pertinent pour estimer la consommation d'énergie de décrire les situations d'habitation puis d'y adjoindre une description des pratiques domestiques.

Notre objectif est de construire une modélisation de la consommation d'énergie domestique (CED) intégrant deux familles de déterminants à savoir les caractéristiques physiques des logements (surface, qualité de l'isolation) ainsi que les pratiques domestiques. Dans cette partie nous commencerons notre travail en présentant les données PHEBUS que nous utiliserons dans ce chapitre en raison de la qualité supérieure des données de consommation d'énergie, par rapport au données ENERGIHAB qui étaient plus riches pour décrire les pratiques domestiques (Partie 1). Nous reviendrons ensuite sur les modèles de régression linéaire qui sont très répandus dans la littérature pour quantifier les effets des deux familles de déterminants qui nous intéressent (Partie 2). Nous explorerons plusieurs limites de ce type de modélisation. Nous étudierons notamment la sensibilité de cette modélisation – et en particulier la sélection des variables et la valeur des effets associés - à l'indicateur de consommation choisi (énergie finale, énergie finale par personne et énergie finale par mètre carré du logement). Aussi, nous étudierons la sensibilité des résultats à l'échantillon considéré, en restreignant l'étude tantôt aux appartements, tantôt aux maisons individuelles etc. A partir de ces calculs et en les croisant avec la revue de littérature effectuée au chapitre 1, et les calculs menés au long du chapitre 2, nous montrerons l'intérêt de réaliser un modèle qui considère à la fois une classification des situations d'habitation pour différencier les types de logements et les types de ménages qui les habitent, et aussi une régression sur les paramètres extensifs (surface, qualité de l'isolation etc.). Nous proposerons un modèle de mélange de régressions multilinéaires à proportions logistiques pour remplir cet objectif. Le modèle effectue alors pendant son entraînement le calcul simultané d'une classification des situations d'habitation ainsi qu'une régression sur chacun des classes. Les pratiques domestiques seront modélisées à l'aide de variables synthétiques, selon la méthode décrite au chapitre 2. Etant liées aux contextes résidentiels, elles serviront alors de variables supplémentaires pour analyser la typologie des situations d'habitation ainsi construite. L'idée de l'algorithme, son entraînement, son évaluation et l'analyse des résultats seront présentés dans la partie 3.

1. Présentation des données

1.1 Les données issues de l'enquête PHEBUS

Nous mobilisons un ensemble de 29 variables décrivant les ménages, les logements, les comportements domestiques et les consommations d'énergie. Pour aller plus loin dans la compréhension des données et des modèles, nous proposons de ventiler la présentation statistique par type d'habitat (Tableau 17). L'analyse du tableau de données, en relation avec les données INSEE permet de voir que le l'ensemble des données après filtrage n'est pas représentatif des ménages et des logements français de 2013. Par exemple, on observe qu'il contient 27% d'appartements contre 35% en moyenne nationale. Par ailleurs le nombre de propriétaires est largement surestimé puisqu'ils représentent ici près de 75% des données contre 58% selon l'INSEE.

Nous avons choisi pour la présentation statistique des données de représenter les données par type de logement. Cette représentation permet d'observer des différences en termes d'occupation qui seront utiles lors des analyses ultérieures. En particulier, on observe que les quintiles de revenus inférieurs et les personnes seules sont sur-représentés dans le sous-échantillon composé des appartements contrairement aux maisons individuelles. Ainsi qu'attendu, les appartements présentent des surfaces inférieures à la moyenne de l'échantillon (64 m² contre 98 m²), sont plus fréquemment localisés dans les zones urbaines ou péri-urbaines et bénéficient plus fréquemment de chauffages collectifs que les maisons. Les performances thermiques des logements sont modélisées par le Diagnostic de Performance Energétique (DPE) qui estime le niveau de consommation (théorique) par mètre carré du logement nécessaire au chauffage du logement à 19°C. En moyenne, on observe que les DPE sont similaires pour toutes les catégories de logement. En termes de consommation énergétique, les appartements présentent en moyenne des consommations énergétiques en énergie finale (FEC) inférieures de 30% à la moyenne de l'échantillon mais qui est relativement proche de la moyenne en termes de consommation ramenée à la taille du logement (FECM2) ou à la taille du ménage (FECP).

Tableau 17 : Résumé statistique des variables extraites de la base de données de l'enquête PHEBUS. Les données sont également présentées par catégorie de logement à des fins pédagogiques. Pour les variables catégorielles, les proportions des modalités sont précisées en % entre parenthèses. Pour les variables numériques, les écarts types sont donnés entre parenthèses. PR : Personne de référence du ménage. Source : Traitements de l'auteur à partir de la base de données PHEBUS.

Variable	Population N = 2291	Appartements N = 626	Maisons multifamiliales N = 45	Maisons individuelles N = 1170	Maisons mitoyennes N = 450
Caractéristiques du ménage					
Date d'arrivée dans le logement	1995 (15)	2000 (11)	1999 (12)	1992 (15)	1995 (15)
Nombre de personnes	2,6 (1,3)	2,3 (1,4)	2,9 (1,5)	2,7 (1,3)	2,6 (1,3)
Quintile de revenus					
Pas de réponse	162 (7,1%)	46 (7,3%)	2 (4,4%)	86 (7,4%)	28 (6,2%)
Q1	420 (18%)	156 (25%)	5 (11%)	162 (14%)	97 (22%)
Q2	444 (19%)	149 (24%)	10 (22%)	194 (17%)	91 (20%)
Q3	428 (19%)	108 (17%)	10 (22%)	231 (20%)	79 (18%)
Q4	415 (18%)	91 (15%)	8 (18%)	227 (19%)	89 (20%)
Q5	422 (18%)	76 (12%)	10 (22%)	270 (23%)	66 (15%)
Composition du ménage					
Couple avec enfants	714 (31%)	128 (20%)	7 (16%)	447 (38%)	132 (29%)
Couple sans enfant	842 (37%)	171 (27%)	23 (51%)	473 (40%)	175 (39%)
Famille monoparentale	187 (8,2%)	77 (12%)	3 (6,7%)	55 (4,7%)	52 (12%)
Plusieurs personnes (pas de famille)	42 (1,8%)	16 (2,6%)	0 (0%)	22 (1,9%)	4 (0,9%)
Personne seule	506 (22%)	234 (37%)	12 (27%)	173 (15%)	87 (19%)
Catégorie socio-professionnelle					
Artisans & Gérants d'entreprise	154 (6,7%)	29 (4,6%)	1 (2,2%)	93 (7,9%)	31 (6,9%)
Employés	520 (23%)	183 (29%)	13 (29%)	232 (20%)	92 (20%)
Cadres	416 (18%)	115 (18%)	9 (20%)	220 (19%)	72 (16%)
Agriculteurs	49 (2,1%)	1 (0,2%)	0 (0%)	41 (3,5%)	7 (1,6%)
Professions intermédiaires	531 (23%)	136 (22%)	13 (29%)	266 (23%)	116 (26%)
Pas de profession	24 (1,0%)	11 (1,8%)	0 (0%)	8 (0,7%)	5 (1,1%)
Ouvriers	597 (26%)	151 (24%)	9 (20%)	310 (26%)	127 (28%)
Statut d'activité de la PR					
Active	1200 (52%)	349 (56%)	25 (56%)	576 (49%)	250 (56%)
Inactive	215 (9,4%)	103 (16%)	7 (16%)	65 (5,6%)	40 (8,9%)
Retraitée	876 (38%)	174 (28%)	13 (29%)	529 (45%)	160 (36%)
Age de la PR	55,4 (15,2)	51,4 (16,1)	52,1 (14,3)	58,0 (13,9)	54,7 (15,8)
Caractéristiques du logement					
Nombre moyen de pièces du logement	4,3 (1,5)	3,2 (1,1)	4,4 (1,2)	4,9 (1,3)	4,5 (1,4)
Surface du logement (m^2)	98,0 (43,9)	64,0 (23,6)	90,6 (29,7)	115,8 (42,9)	100,0 (41,3)
Statut d'occupation du logement					
Autre	74 (3,2%)	22 (3,5%)	1 (2,2%)	29 (2,5%)	22 (4,9%)
Propriétaires et accédants	1,711 (75%)	275 (44%)	33 (73%)	1,073 (92%)	330 (73%)
Locataires du secteur privé	264 (12%)	148 (24%)	5 (11%)	57 (4,9%)	54 (12%)
Locataires du secteur public	242 (11%)	181 (29%)	6 (13%)	11 (0,9%)	44 (9,8%)
Date de construction du logement					
<1919	351 (15%)	71 (11%)	2 (4,4%)	154 (13%)	124 (28%)
1919-1945	189 (8,2%)	41 (6,5%)	2 (4,4%)	68 (5,8%)	78 (17%)

1946-1970	456 (20%)	183 (29%)	8 (18%)	182 (16%)	83 (18%)
1971-1990	768 (34%)	202 (32%)	21 (47%)	439 (38%)	106 (24%)
1991-2005	381 (17%)	96 (15%)	7 (16%)	236 (20%)	42 (9,3%)
>2005	146 (6,4%)	33 (5,3%)	5 (11%)	91 (7,8%)	17 (3,8%)
Zone urbaine (nombre d'habitants)					
<5k	785 (34%)	61 (9,7%)	6 (13%)	578 (49%)	140 (31%)
5k-20k	226 (9,9%)	25 (4,0%)	6 (13%)	146 (12%)	49 (11%)
20k-100k	290 (13%)	74 (12%)	5 (11%)	139 (12%)	72 (16%)
100k-200k	123 (5,4%)	43 (6,9%)	3 (6,7%)	38 (3,2%)	39 (8,7%)
200k-2M	558 (24%)	211 (34%)	15 (33%)	210 (18%)	122 (27%)
Paris	309 (13%)	212 (34%)	10 (22%)	59 (5,0%)	28 (6,2%)
Type de chauffage principal					
Chauffage urbain	32 (1,4%)	32 (5,1%)	0 (0%)	0 (0%)	0 (0%)
Electrique individuel	611 (27%)	197 (31%)	18 (40%)	295 (25%)	101 (22%)
Fioul collectif	29 (1,3%)	29 (4,6%)	0 (0%)	0 (0%)	0 (0%)
Fioul individuel	305 (13%)	8 (1,3%)	1 (2,2%)	254 (22%)	42 (9,3%)
Gaz collectif	101 (4,4%)	100 (16%)	0 (0%)	1 (<0,1%)	0 (0%)
Gaz individuel	750 (33%)	187 (30%)	24 (53%)	295 (25%)	244 (54%)
Pompe à chaleur	104 (4,5%)	7 (1,1%)	0 (0%)	82 (7,0%)	15 (3,3%)
Bois	208 (9,1%)	2 (0,3%)	1 (2,2%)	174 (15%)	31 (6,9%)
Autre	151 (6,6%)	64 (10%)	1 (2,2%)	69 (5,9%)	17 (3,8%)
Chauffage individuel ou collectif					
Collectif	235 (10%)	221 (35%)	0 (0%)	10 (0,9%)	4 (0,9%)
Individuel	2,056 (90%)	405 (65%)	45 (100%)	1,160 (99%)	446 (99%)
Diagnostic de performance énergétique (DPE en kWh/m²)	282 (149)	291 (152)	220 (86)	277 (146)	288 (153)
Travaux de rénovations depuis 2008					
Aucun	1,137 (50%)	372 (59%)	19 (42%)	537 (46%)	209 (46%)
Changement d'équipement	264 (12%)	60 (9,6%)	4 (8,9%)	155 (13%)	45 (10%)
Isolation thermique	380 (17%)	87 (14%)	7 (16%)	199 (17%)	87 (19%)
Isolation thermique et changement d'équipement	275 (12%)	29 (4,6%)	10 (22%)	164 (14%)	72 (16%)
Autre	235 (10%)	78 (12%)	5 (11%)	115 (9,8%)	37 (8,2%)
Degré jour unifiés (DJU)					
1299	147 (6,4%)	62 (9,9%)	8 (18%)	52 (4,4%)	25 (5,6%)
1841	250 (11%)	35 (5,6%)	2 (4,4%)	177 (15%)	36 (8,0%)
1896	181 (7,9%)	31 (5,0%)	2 (4,4%)	114 (9,7%)	34 (7,6%)
1953	266 (12%)	35 (5,6%)	3 (6,7%)	151 (13%)	77 (17%)
1964	73 (3,2%)	24 (3,8%)	4 (8,9%)	31 (2,6%)	14 (3,1%)
2123	699 (31%)	257 (41%)	18 (40%)	270 (23%)	154 (34%)
2283	383 (17%)	111 (18%)	7 (16%)	212 (18%)	53 (12%)
2360	292 (13%)	71 (11%)	1 (2,2%)	163 (14%)	57 (13%)
Caractéristiques de consommation d'énergie finale					
Consommation en énergie finale totale (FEC)	18 847 (12 334)	13 518 (10 678)	16 187 (9 869)	21 607 (12 361)	19,348 (12 226)
Consommation en énergie finale totale par m² de surface (FECM2)	203 (125)	218 (156)	177 (80)	197 (114)	199 (108)
Consommation en énergie finale totale par habitant (FECP)	8 979 (7 231)	7,335 (6 416)	7 273 (5 482)	9 879 (7 438)	9 098 (7 496)

1.2 Etude des corrélations

Pour compléter la présentation des données nous proposons d'étudier les corrélations entre les variables à l'aide d'un graphique (Figure 39). En effet, l'analyse simultanée des corrélations entre un grand nombre de variables n'est pas aisée. Nous nous appuyons sur la méthode proposée par la fonction *network_plot* du package *corrr*[27]. Cette visualisation suppose premièrement de créer un point pour représenter chaque variable. Ensuite, un lien entre chaque point est créé et son apparence est fixée en fonction du niveau de corrélation entre ces deux variables : plus la corrélation est forte, plus le lien est épais et moins il est transparent. De cette manière, deux variables très peu liées seront liées par un trait fin et transparent. Un code couleur permet de distinguer les corrélations négatives (rouge) et les corrélations positives (vert). Aussi, on choisit un seuil de corrélation minimal afin de ne pas avoir une figure surchargée : dans notre cas nous choisissons un seuil de 0,25. Finalement la position des points dans l'espace de représentation est calculée à l'aide d'un algorithme appelé « Multidimensional scaling » (MDS), appliqué à la valeur absolue des corrélations[28]. Cette visualisation facilite la visualisation des interrelations linéaires entre les variables une à une mais aussi entre des groupes de variables. En centrant l'analyse de la figure sur les 3 indicateurs de consommations, on observe 4 groupes de variables :

- Un groupe autour de l'indicateur FECP à gauche, avec les variables caractérisant la composition du ménage, son revenu et son cycle de vie, ainsi que des variables caractérisant l'occupation du logement.
- Un groupe autour de l'indicateur FECM2, à l'opposé, avec les variables caractérisant le système de chauffage d'une part et des variables de comportements thermique d'autre part.
- Un groupe de variables presque autonome, en haut, avec des variables caractérisant le comportements de régulations du ménage.
- Un dernier groupe, au centre, autour de la variable FEC, qui se trouve liée particulièrement avec des variables telles que la surface du logement et le nombre d'équipements mais aussi avec des variables des autres groupes telles que la composition du ménage et le système de chauffage.

Cette analyse en « pôles » permet de rendre compte du lien entre les indicateurs de consommations, mais aussi des différents liens que ces indicateurs ont avec les variables explicatives utilisées classiquement dans les modèles de CED. L'analyse souligne l'intérêt de croiser les approches de modélisation des trois indicateurs afin de saisir au mieux les dynamiques que chacun traduit :

---

[27] Une présentation du package est accessible au lien suivant : https://cran.r-project.org/web/packages/corrr/index.html
La version utilisée est la version 0.4.4.

[28] Le principe de l'algorithme MDS est de calculer (construire) un espace abstrait de petite dimension (généralement 2 ou 3) pour représenter des données originales qui sont décrites en grande dimension. Le critère utilisé permet de minimiser la distorsion des distances relatives entre les individus dans l'espace final. Dans notre cas, la distance utilisée dans l'espace initial est la corrélation entre deux variables et la distance dans l'espace final est une distance géométrique en dimension 2.

- la composition du ménage, l'occupation du logement et les comportements de régulation seraient mieux saisis avec FECP,
- la demande en chauffage et le système de chauffage seraient mieux saisis par l'indicateur FECM2
- la taille du logement, la composition du ménage, le niveau d'équipement seraient particulièrement bien saisis par l'indicateur FEC.

Cette analyse préliminaire ouvre la voie à une première modélisation linéaire de ces trois indicateurs qui permettra de confirmer les dépendances identifiées.

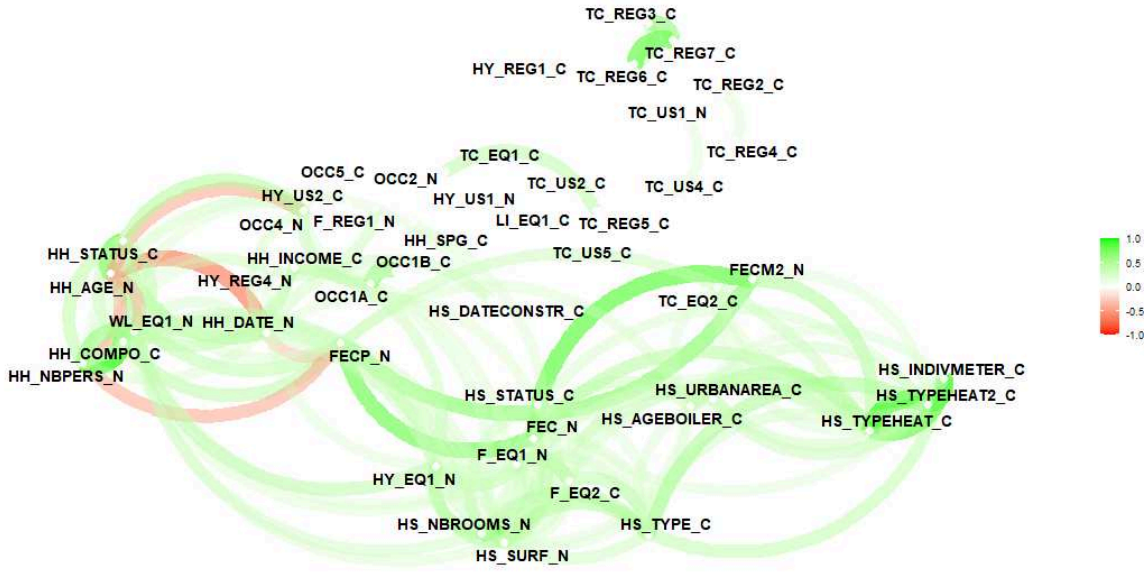


Figure 39: Graphe décrivant le degré de liaison des variables deux à deux. L'épaisseur et la couleur des liens entre chaque variable est lié à la force de la liaison. La liaison est mesurée à l'aide du V de Cramer ajusté entre deux variables catégorielles, du facteur de corrélation de Spearman entre deux variables numériques et par ANOVA entre une variable numérique et une variable catégorielle. Seuls les coefficients supérieurs à 0,25 sont représentés. Source : Traitements de l'auteur à partir des données PHEBUS.

2. Modélisation de référence : un modèle linéaire de la CED

2.1.1 Régression linéaire simple entre la CED et les facteurs explicatifs

Dans ce paragraphe, on fait l'hypothèse qu'il existe un lien linéaire entre chacune des variables et les logarithmes des 3 indicateurs de consommations d'énergie finale (énergie totale de symbole FEC exprimée en kWh, énergie totale par personne de symbole FECP exprimée en kWh/p et énergie totale par mètre de surface du logement de symbole FECM2 exprimée en kWh). Cette hypothèse très simple permet dans un premier temps de tester la significativité statistique du lien entre les variables. Dans le cas de variables qualitatives, on ne peut pas faire d'hypothèse de linéarité : on va essayer plutôt d'observer si les moyennes observée des indicateurs sur chacune des modalités sont « assez » différentes

pour estimer s'il existe un lien statistique entre la variable et l'indicatrice de consommation : on parle alors plutôt d'analyse de la variance à un facteur. Sur le plan méthodologique, nous avons remplacé les variables DPE, SURF et SURFP par les logarithmes de ces variables. Les résultats sont synthétisés dans le tableau suivant (Tableau 18).

Tableau 18 : Synthèse des effets individuels des variables explicatives par rapport aux trois variables expliquées (FEC, FECM2, et FECP). Pour les variables quantitatives l'effet estimé est le coefficient de régression, pour les variables qualitatives il s'agit de l'écart moyen de la variable expliquée par rapport à la modalité de référence. La p-valeur exprime la probabilité de se tromper en considérant la non-nullité de l'effet. Un code couleur permet de repérer les p-valeurs inférieures à 10%. Source : Auteur d'après des calculs effectués sur la base PHEBUS).

	FEC		FECM2		FECP	
	Effet estimé	P-valeur	Effet estimé	P-valeur	Effet estimé	P-valeur
HH_DATE_N	-0,004	0%	-0,002	0%	-0,008	0%
HH_NBPERS_N	0,051	0%	0,010	3%	-0,119	0%
HH_INCOME_C (réf: Revenu non connu)						
Q1	0,001	99%	0,087	0%	0,090	1%
Q2	-0,022	48%	0,045	10%	0,020	54%
Q3	0,032	29%	0,060	3%	-0,014	67%
Q4	0,041	18%	0,039	15%	-0,051	12%
Q5	0,109	0%	0,037	17%	-0,019	58%
HH_COMPO_C (réf: Couple sans enfant)						
Couple avec enfants	0,023	14%	0,014	32%	-0,256	0%
Famille monoparentale	-0,075	0%	-0,007	77%	-0,180	0%
Plusieurs personnes	-0,023	67%	-0,005	92%	-0,123	2%
Personne seule	-0,195	0%	-0,032	6%	0,106	0%
HH_SPG_C (Réf : Artisans et Chefs d'entreprise)						
Employés	-0,116	0%	0,023	39%	-0,028	40%
Cadres	-0,055	7%	-0,020	45%	-0,063	6%
Agriculteurs	0,162	0%	0,092	5%	0,169	0%
Professions intermédiaires	-0,088	0%	-0,012	64%	-0,058	8%
Pas de profession	-0,132	9%	0,013	85%	0,045	60%
Ouvriers	-0,061	4%	0,050	6%	-0,055	9%
HH_STATUS_C (Réf : Actif)						
Inactif	-0,081	0%	-0,013	56%	0,011	68%
Retraité	0,027	6%	0,008	56%	0,228	0%
HH_AGE_N	0,003	0%	0,001	6%	0,009	0%
HS_TYPE_C (Réf : Appartement)						
Maison multifamiliale	0,167	0%	0,006	90%	0,034	54%
Maison individuelle	0,262	0%	0,004	76%	0,155	0%
Maison mitoyenne	0,204	0%	0,015	39%	0,108	0%
HS_NBROOMS_N	0,084	0%	-0,013	0%	0,033	0%
HS_SURF_N (log)	0,805	0%	-0,195	0%	0,411	0%
HS_SURFP_N (log)	0,188	0%	-0,171	0%	0,829	0%
HS_STATUS_C (Réf: Autres)						
Propriétaire	0,014	72%	-0,060	9%	-0,151	0%
Locataire - secteur privé	-0,174	0%	-0,034	38%	-0,258	0%
Locataire - secteur public	-0,145	0%	-0,038	34%	-0,286	0%
HS_DATECONSTR_C (Réf : <1919)						
1919-1945	-0,101	0%	-0,136	0%	-0,192	0%
1946-1970	-0,026	18%	-0,068	0%	-0,076	0%
1971-1990	-0,012	56%	0,032	6%	-0,012	56%
1991-2005	0,036	6%	0,036	3%	0,062	0%
>2005	-0,016	29%	-0,033	1%	0,006	73%
HS_URBANAREA_C						
5k-20k	-0,081	0%	0,049	0%	-0,047	1%
20k-100k	0,060	0%	0,096	0%	0,067	0%
100k-200k	-0,002	93%	0,028	9%	-0,016	44%
200k-2M	0,006	78%	0,029	11%	0,020	37%
Paris	-0,024	31%	-0,030	16%	-0,020	46%
HS_TYPEHEAT_C						
Electrique individuel	0,521	0%	0,396	0%	0,570	0%
Fioul - collectif	0,894	0%	0,857	0%	0,939	0%
Fioul individuel	0,891	0%	0,624	0%	0,904	0%
Gas collectif	0,787	0%	0,783	0%	0,821	0%
Gaz individuel	0,782	0%	0,593	0%	0,775	0%

Pompe à chaleur	0,609	0%	0,325	0%	0,569	0%
Bois	0,895	0%	0,644	0%	0,839	0%
Autre	0,606	0%	0,456	0%	0,581	0%
HS_TYPEHEAT2_C						
Individuel						
F_EQ1_N	0,061	0%	0,002	76%	0,025	0%
F_EQ2_C (Réf : Congélateur indépendant : Non)						
Congélateur indépendant : Oui	0,151	0%	0,016	20%	0,067	0%
OCC5_C						
Basse	0,074	28%	-0,061	31%	-0,153	4%
Moyenne	0,080	1%	0,021	44%	-0,110	0%
F_REG1_N	-0,002	83%	-0,015	5%	-0,040	0%
HY_EQ1_N	0,078	0%	-0,008	17%	0,013	7%
HY_US1_N	-0,054	0%	-0,017	31%	-0,019	33%
HY_US2_C (Réf : Bain : Non)						
Bain : oui	0,025	12%	-0,012	40%	-0,122	0%
HY_REG1_C : (Gestes de régulation : Non)						
Gestes de régulation : Oui	-0,002	90%	0,013	28%	0,010	49%
HY_REG4_N	0,027	0%	-0,003	63%	-0,043	0%
LI_EQ1_C (Réf : Halogènes : Non)						
Halogènes : oui	0,070	0%	0,015	23%	0,071	0%
TC_EQ1_C (Chauffage aux. : non)						
Chauffage auxiliaire oui	0,002	95%	0,033	13%	0,046	9%
TC_EQ2_C (Réf : Régulation centralisée et non prog.)						
Régulation centralisée et programmable	-0,006	82%	-0,017	45%	-0,042	13%
Régulation décentralisée	-0,039	12%	-0,009	70%	-0,050	8%
Pas d'éq. De régulation	-0,109	0%	0,002	95%	-0,091	0%
TC_US1_N	0,022	0%	0,024	0%	0,013	2%
TC_US2_C : Aucun geste de régulation du chauffage						
Jamais	-0,018	83%	-0,120	10%	0,137	12%
Occasionnellement	-0,007	75%	-0,018	34%	-0,013	56%
Souvent	0,008	62%	-0,023	9%	-0,003	85%
TC_US4_C						
Ouvre les fenêtres pour réguler la température	0,044	21%	0,104	0%	0,004	91%
TC_US5_N	0,064	0%	0,046	0%	0,062	0%
TC_REG2_C (Réf : Eteint régulièrement le chauffage : Oui)						
Jamais	0,031	5%	0,031	3%	0,015	37%
Parfois	-0,028	18%	-0,032	9%	-0,043	6%
TC_REG3_C (Réf : Ne chauffe pas certaines pièces : Non)						
Ne chauffe pas certaines pièces : Oui	-0,068	0%	-0,052	1%	0,004	88%
TC_REG4_C (Réf : Coupe le chauffage pendant l'aération)						
Non	0,031	3%	0,030	2%	0,037	2%
Parfois	-0,004	90%	-0,013	64%	-0,013	71%
TC_REG5_C (Réf : Eteint le chauffage pendant les vacances)						
Rien	-0,040	2%	0,031	4%	-0,044	2%
Eteint le chauffage	-0,175	0%	-0,068	0%	-0,128	0%
TC_REG6_C (Réf : Réduit le chauffage par restriction: Non)						
Réduit le chauffage par restriction: Oui	-0,084	0%	-0,023	19%	-0,064	0%
TC_REG7_C (Réf : Ne chauffe pas certaines pièces par restriction : Non)						
Ne chauffe pas certaines pièces par restriction : Oui	-0,066	0%	-0,020	17%	-0,043	2%
HS_INDIVMETER_C (Réf : Compteur individuel : Non)						
Compteur individuel : oui	0,013	59%	-0,131	0%	0,009	73%
HS_AGEBOILER_C (Réf: <1991)						
1991-2001	-0,211	0%	-0,122	0%	-0,219	0%
2001-2012	-0,083	0%	-0,045	0%	-0,052	0%
Pas de chauffage individuel	-0,041	1%	-0,028	4%	-0,019	25%
WL_EQ1_N	0,016	0%	0,000	89%	-0,026	0%
OCC1A_C (Réf : >12h)						
<4h	0,149	2%	0,031	60%	-0,072	32%
4-8h	0,039	56%	-0,001	98%	-0,104	16%
8-12h	-0,083	27%	-0,035	61%	-0,178	3%
OCC1B_C (Réf : >12h)						
<4h	0,170	1%	0,024	69%	0,009	90%
4-8h	0,065	35%	-0,017	78%	-0,032	67%
8-12h	0,043	58%	0,041	56%	0,040	64%
OCC2_N	-0,002	0%	-0,002	0%	0,002	1%
OCC4_N	-0,012	0%	-0,006	0%	-0,013	0%
DPE_N (log)	0,002	0%	0,002	0%	0,002	0%
DJU_N	0,000	0%	0,000	0%	0,000	0%

L'analyse variable par variable permet de faire plusieurs remarques. En premier lieu, toutes les variables sélectionnées sont liées statistiquement à au moins une des trois variables expliquées. Toutefois, selon cette analyse certaines variables semblent être liées à la consommation totale mais non à la consommation par unité de surface ou du moins avec des effets différents. Parmi ces variables on cite la composition du ménage, la catégorie socio-professionnelle, le statut d'activité, le statut d'occupation du logement. Une seconde remarque est la dépendance de certains coefficients à l'indicateur de consommation considéré. Contrairement à la variable FEC, on observe que les valeurs prises par la variable FECM2 diminue tendanciellement lorsque la surface augmente. Cette relation négative est intéressante car elle illustre le fait qu'à logement et ménage donnés, l'augmentation en nombre ou en intensité des services énergétiques permise par l'augmentation de la surface n'est en moyenne par proportionnel à la surface qui serait « ajoutée ».

Aussi, l'indicateur FECM2 est plutôt lié aux variables décrivant l'intensité du chauffage et sa régulation ainsi que les périodes de d'inoccupation longues. On remarque enfin que les indicateurs FECM2 et FECP sont tous deux plus liés aux revenus faibles que l'indicateur FEC.

La régression linéaire simple permet de rendre compte de la qualité d'une éventuelle relation linéaire entre les variables explicatives et les indicatrices de consommation d'énergie. Toutefois, ces modèles peinent à rendre compte quantitativement des effets de chacun des phénomènes sous-jacents en raison du croisement des effets des différentes variables (par exemple la taille du ménage et la taille du logement). Le calcul d'une régression multiple permet dans ce sens de créer un modèle explicatif plus fidèle à la réalité où les effets estimés sont calculés « toutes choses égales par ailleurs » afin de peser les effets relatifs.

2.1.2 Une modélisation multilinéaire de la consommation en énergie finale

Le modèle multilinéaire est un modèle de référence dans le champ de la consommation d'énergie domestique. Mathématiquement, ce modèle permet d'estimer la consommation d'énergie d'un ménage comme la somme pondérée « d'effets moyens » de variables supposées indépendantes : les modèles postulent puis calculent un effet de la surface, du nombre de personnes composant le ménage, de la présence d'un équipement ou de la réalisation d'un comportement. Cet effet est calculé par apprentissage statistique sous cinq hypothèses :

- **Hypothèse de linéarité** : on suppose que le lien entre la consommation d'énergie et les variables explicatives identifiées est linéaire.
- **Absence de multi colinéarité** : les variables explicatives ne sont pas (ou peu) corrélées entre elles.
- **Hypothèses de normalité et de biais nul des résidus** : le terme d'erreur (les « résidus ») a une distribution gaussienne, centrée sur 0.
- **Indépendance des résidus** : les résidus sont indépendants des variables explicatives.

Dans ce paragraphe nous calculons les modèles de régression linéaire multiple pour les 3 indicateurs de consommation d'énergie. La méthodologie est la suivante :

- Nous construisons un modèle à partir de l'ensemble des variables.
- Nous analysons les variables indiquées comme statistiquement significatives, ainsi que les effets calculés. En particulier, le critère d'inflation de la variance (Variance Inflation Factor – VIF[29]) est utilisé pour identifier les variables très corrélées entre elles et dont les effets calculés sont entachés d'une forte incertitude. La normalité des résidus est également évaluée.
- Nous calculons ensuite le modèle réduit à l'aide d'un algorithme de sélection de variable selon le critère AIC (Critère d'information d'Akaike), et nous répétons le processus d'analyse des variables et des résidus.
- Pour caractériser la performance d'estimation des modèles nous les entrainons et testons 50 fois (les coefficients de la régressions sont recalculés) sur des échantillons aléatoires. Les indicateurs de performance sont les suivants :

L'erreur absolue moyenne MAE (Mean Absolute Error)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

L'erreur absolue moyenne exprimée en pourcentage MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}$$

L'erreur quadratique moyenne RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Le coefficient de détermination (ou pourcentage de variance expliquée) ou R^2

$$R^2 = 1 - \sum_{i=1}^{N}\frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

Où y_i et $\hat{y}_i$ désignent respectivement le niveau de la variable à prédire et le niveau estimé de cette même variable estimé à la ligne i. $\bar{y}$ désigne la moyenne de la variable y.

[29] Le critère d'inflation de la variance est un indicateur utilisé pour détecter une multi-colinéarité dans le modèle de régression.

Remarque importante : Dans ce travail, on effectue l'ensemble des modélisations en manipulant le logarithme décimal des consommations d'énergie. Toutefois, la performance d'estimation est calculée à partir des consommations en kWh.

Nous présentons ici les modèles sélectionnés par cette procédure pour les 3 indicateurs FEC, FECP et FECM2.

La régression de la consommation totale (FEC)

Le premier modèle réalisé est la régression multilinéaire de la consommation totale (FEC). Une liste des variables sélectionnées dans le modèle final est donnée dans le tableau suivant (Tableau 19). On remarque l'importance des variables suivantes : type de chauffage, surface, DPE, composition du ménage, type de logement, âge du chauffage. Les variables de comportements semblent toutefois jouer un rôle secondaire (en raison d'effets plus petits en valeur absolue et significativité des coefficients limitée).

En termes de précision, ce type de modèle offre des estimations avec une erreur quadratique de 9MWh en moyenne et une erreur absolue moyenne de 6MWh. Les performances moyennes sont indiquées dans le second tableau ci-dessous. Ces erreurs sont très importantes et témoigne d'une difficulté du modèle à modéliser la variabilité des consommations.

Tableau 19 : Liste des variables et des effets calculés pour le modèle multilinéaire de l'indicateur FEC. Source : Auteur après calculs sur la base PHEBUS.

	Terme	Effet estimé	Incertitude	P-valeur
HS_TYPEHEAT_C	Electrique individuel	0,68	0,06	0%
	Fioul - collectif	0,84	0,07	0%
	Fioul individuel	0,84	0,06	0%
	Gas collectif	0,75	0,05	0%
	Gaz individuel	0,84	0,06	0%
	Pompe à chaleur	0,62	0,06	0%
	Bois	0,89	0,06	0%
	Autre	0,65	0,06	0%
HS_SURF_N		0,56	0,04	0%
DPE_N		0,30	0,03	0%
HS_TYPEHEAT2_C	Individuel	-0,28	0,04	0%
HH_COMPO_C (réf: Couple sans enfant)	Couple avec enfants	0,04	0,01	1%
	Famille monoparentale	0,00	0,02	93%
	Plusieurs personnes	0,01	0,04	71%
	Personne seule	-0,05	0,02	0%
TC_US5_N		0,02	0,00	0%
HS_TYPE_C (Réf: Appartement)	Maison multifamiliale	0,09	0,04	1%
	Maison individuelle	0,07	0,02	0%
	Maison mitoyenne	0,04	0,02	6%
LI_EQ1_C	Halogènes: oui	0,04	0,01	0%
HS_AGEBOILER_C (Réf: <1991)	1991-2001	-0,07	0,02	0%
	2001-2012	-0,03	0,02	5%
	NO_INDBOILER	-0,02	0,01	6%
F_EQ1_N		0,01	0,01	1%
TC_REG5_C	Rien	0,00	0,01	96%
	Eteint le chauffage	-0,05	0,02	1%
TC_US1_N		0,01	0,00	2%
HS_DATECONSTR_C	1919-1945	-0,03	0,02	7%
	1946-1970	0,02	0,01	9%
	1971-1990	-0,01	0,01	62%
	1991-2005	0,02	0,01	20%
	>2005	-0,01	0,01	32%
WL_EQ1_N		0,00	0,00	2%
HH_DATE_N		0,00	0,00	7%
DJU_N		0,00	0,00	2%
TC_REG4_C	Non	0,02	0,01	12%
	Parfois	-0,03	0,02	15%
TC_US4_C	Ouvre les fenêtres pour réguler la température	0,05	0,03	7%
OCC4_N		0,00	0,00	7%
F_REG1_N		-0,01	0,01	10%
HY_EQ1_N		0,01	0,01	14%

R^2	RMSE	MAE	MAPE
44,5% [4%]	9089 kWh [614 kWh]	6093 kWh [319 kWh]	32,2% [1,1%]

Il est aussi intéressant d'étudier les corrélations entre les variables au sein de ce modèle et les incertitudes qu'elles génèrent sur l'estimation des coefficients. Le coefficient VIF permet de mesurer cette multi colinéarité. En l'occurrence, les variables caractérisant la surface, le chauffage individuel, le type de logement et l'âge du chauffage ont des coefficients VIF entre 2 et 5. Le VIF pour le type de chauffage est lui de 28, ce qui témoigne d'une colinéarité très forte et d'une incertitude importante sur la valeur du

coefficient calculé. La variable reste toutefois très importante et ne peut être écartée du modèle au risque de dégrader significativement sa performance d'estimation.

L'analyse des résidus montre quant à elle que ces derniers sont plus importants pour les valeurs extrêmes de l'indicateur FEC.

La régression de FECM2

Le Tableau 20 recense les variables sélectionnées pour construire le modèle de l'indicateur FECM2. Beaucoup de variables sont communes avec le modèle précédent. La différence notable avec le modèle de l'indicateur FEC est la sélection de la variable modélisant la surface par personne en plus de la variable de surface. On notera d'ailleurs que les effets calculés sont négatifs : la valeur de l'indicateur FECM2 décroit tendanciellement avec la surface et la surface par personne. Par ailleurs quelques variables modélisant l'effet de comportements (régulation thermique, niveau d'équipement) sont sélectionnées.

En termes de VIF, nous observons que la date de construction, la zone urbaine, et le chauffage individuel ont un VIF entre 2 et 5, tandis que d'autres ont des VIF très élevés : le type de chauffage (32), la surface par personne (16), nombre de personnes (14), surface (12).

En termes de performance d'estimation, le modèle FECM2 explique en moyenne 42% de la variance et offre une erreur quadratique de 94 kWh/m².

Tableau 20 : Liste des variables et des effets calculés pour le modèle multilinéaire de l'indicateur FECM2. Source : Auteur après calculs sur la base PHEBUS.

	Terme	Effet estimé	Incertitude	P-valeur
	(Intercept)	2,75	0,88	0%
HS_TYPEHEAT_C	Electrique individuel	0,67	0,06	0%
	Fioul - collectif	0,84	0,07	0%
	Fioul individuel	0,83	0,06	0%
	Gas collectif	0,74	0,05	0%
	Gaz individuel	0,84	0,06	0%
	Pompe à chaleur	0,62	0,06	0%
	Bois	0,89	0,06	0%
	Autre	0,64	0,06	0%
DPE_N		0,29	0,03	0%
HS_SURFP_N		-0,23	0,08	0%
HS_TYPEHEAT2_C	Individuel	-0,27	0,04	0%
TC_US5_N		0,02	0,00	0%
LI_EQ1_C (Réf : Halogènes: Non)	Halogènes: oui	0,04	0,01	0%
HH_NBPERS_N		-0,02	0,01	26%
HS_TYPE_C (Réf: Appartement)	Maison multifamiliale	0,09	0,04	1%
	Maison individuelle	0,08	0,02	0%
	Maison mitoyenne	0,04	0,02	2%
HS_AGEBOILER_C (Réf: <1991)	1991-2001	-0,07	0,02	0%
	2001-2012	-0,03	0,02	7%
	Pas de chauffage individuel	-0,02	0,01	5%
F_EQ1_N		0,02	0,01	0%
TC_REG5_C	Rien	0,00	0,01	98%
	Eteint le chauffage	-0,05	0,02	1%
TC_US1_N		0,01	0,00	2%
HS_DATECONSTR_C (Réf: <1919)	1919-1945	-0,03	0,02	8%
	1946-1970	0,02	0,02	15%
	1971-1990	-0,01	0,01	62%
	1991-2005	0,02	0,01	20%
	>2005	-0,01	0,01	34%
HS_URBANAREA_C (Réf: <5k)	5k-20k	0,02	0,01	26%
	20k-100k	0,04	0,02	2%
	100k-200k	0,01	0,01	34%
	200k-2M	0,00	0,02	78%
	Paris	-0,02	0,02	23%
TC_REG4_C	Non	0,02	0,01	13%
	Parfois	-0,03	0,02	14%
TC_US4_C	Ouvre les fenêtres pour réguler la température	0,05	0,03	8%
OCC4_N		0,00	0,00	5%
DJU_N		0,00	0,00	5%
WL_EQ1_N		0,00	0,00	10%
HS_SURF_N		-0,21	0,09	2%
HH_DATE_N		0,00	0,00	8%
F_REG1_N		-0,01	0,01	10%
HY_EQ1_N		0,01	0,01	10%
HY_US1_N		0,02	0,01	23%

R^2	RMSE	MAE	MAPE
42% [5%]	94 kWh/m² [7 kWh/m²]	64 kWh/m² [3%]	32% [1%]

La régression de FECP

Le dernier indicateur est FECP : son modèle est présenté dans le Tableau 21. On observe peu de différences dans les variables sélectionnées par rapport au modèle de l'indicateur FEC. Toutefois, les performances d'estimation sont plutôt élevées : le modèle modélise 59% de la variance et présente une erreur quadratique de 4,5 MWh/p.

Tableau 21 : Liste des variables et des effets calculés pour le modèle multilinéaire de l'indicateur FECP. Source : Auteur après calculs sur la base PHEBUS.

	Terme	Effet estimé	Incertitude	P-valeur
	(Intercept)	3,79	0,91	0%
HS_TYPEHEAT_C	Electrique individuel	0,65	0,06	0%
	Fioul - collectif	0,86	0,07	0%
	Fioul individuel	0,81	0,06	0%
	Gas collectif	0,75	0,06	0%
	Gaz individuel	0,81	0,06	0%
	Pompe à chaleur	0,60	0,06	0%
	Bois	0,86	0,06	0%
	Autre	0,62	0,06	0%
HS_SURF_N		0,49	0,04	0%
DPE_N		0,29	0,03	0%
HS_TYPEHEAT2_C	Individuel	-0,24	0,04	0%
HH_COMPO_C (Réf: Couple sans enfant)	Couple avec enfants	-0,22	0,01	0%
	Famille monoparentale	-0,10	0,02	0%
	Plusieurs personnes	-0,08	0,04	4%
	Personne seule	0,23	0,02	0%
TC_US5_C		0,03	0,00	0%
HS_TYPE_C (Réf: Appartement)	Maison multifamiliale	0,08	0,04	4%
	Maison individuelle	0,07	0,02	0%
	Maison mitoyenne	0,04	0,02	7%
LI_EQ1_C (Réf : Halogènes: Non)	Halogènes: oui	0,05	0,01	0%
HS_AGEBOILER_C (Réf: <1991)	1991-2001	-0,07	0,02	0%
	2001-2012	-0,03	0,02	10%
	Pas de chauffage individuel	-0,02	0,01	19%
F_EQ1_N		0,02	0,01	1%
TC_REG5_C	Rien	0,00	0,01	89%
	Eteint le chauffage	-0,05	0,02	1%
TC_US1_N		0,01	0,00	1%
HS_DATECONSTR_C	1919-1945	-0,04	0,02	4%
	1946-1970	0,02	0,02	30%
	1971-1990	-0,01	0,02	50%
	1991-2005	0,02	0,01	30%
	>2005	-0,01	0,01	59%
WL_EQ1_N		0,00	0,00	75%
HH_DATE_N		0,00	0,00	0%
DJU_N		0,00	0,00	8%
TC_REG4_C	Non	0,01	0,01	26%
	Parfois	-0,04	0,02	13%
TC_US4_C	Ouvre les fenêtres pour réguler la température	0,04	0,03	13%
OCC4_N		0,00	0,00	73%
F_REG1_N		-0,01	0,01	12%
HY_EQ1_N		0,01	0,01	3%

R^2	RMSE	MAE	MAPE
59% [4%]	4,5 MWh/p [0,3 MWh/p]	2,9 MWh/p [0,2 MWh/p]	32 % [1%]

La modélisation multilinéaire des trois indicateurs permet de mettre en avant le caractère déterminant de facteurs décrivant les contextes résidentiels ainsi que de variables comportementales. Le type de logement, la surface, le degré d'isolation thermique, le type de chauffage, la composition du ménage sont des variables qui interviennent de manière récurrente dans ces modèles. Il est cependant intéressant de remarquer que la variable modélisant les revenus du ménage n'apparait pas dans les sélections des variables explicatives. Si cela n'exclue pas le revenu des ménages de l'explication des consommations d'énergie, cela met en tout cas en avant le fait que son « pouvoir différenciant » est plus faible que les variables précitées.

2.1.3 Une modélisation linéaire de la consommation en énergie finale sur des sous-ensembles

Dans ce paragraphe nous proposons d'étudier la sensibilité du modèle de l'indicateur FEC à l'ensemble de données considéré. Nous présentons les résultats en termes de variables sélectionnées et de performances prédictives dans les tableaux suivants, pour les appartements (Tableau 22), les maisons (Tableau 23).

Tableau 22 : Liste des variables et des effets calculés pour le modèle multilinéaire de l'indicateur FEC pour les appartements uniquement. Les performances d'estimation sont également données. Source : Auteur après calculs sur la base PHEBUS.

	Terme	Effet estimé	Incertitude	P-valeur
	(Intercept)	1,14	0,21	0%
HS_TYPEHEAT_C (Réf: Chauffage urbain)	Electrique individuel	0,78	0,09	0%
	Fioul - collectif	0,80	0,07	0%
	Fioul individuel	1,15	0,13	0%
	Gas collectif	0,71	0,05	0%
	Gaz individuel	1,07	0,09	0%
	Pompe à chaleur	0,75	0,10	0%
	Bois	1,05	0,22	0%
	Autre	0,71	0,06	0%
HH_NBPERS_N		-0,01	0,03	65%
DPE_N		0,47	0,05	0%
HS_SURF_N		0,88	0,15	0%
HS_TYPEHEAT2_C (Réf: Collectif)	Individuel	-0,45	0,07	0%
TC_REG5_C	Rien	0,02	0,03	43%
	Eteint le chauffage	-0,06	0,04	14%
HY_US1_N		0,06	0,02	1%
F_REG1_N		-0,05	0,01	0%
HY_REG4_N		0,04	0,02	1%
HS_SURFP_N		-0,34	0,15	2%
TC_US5_N		0,02	0,01	4%

R^2	RMSE	MAE	MAPE
59% [9%]	6583 kWh [990 kWh]	4285 kWh [557 kWh]	34% [3%]

Tableau 23 : Liste des variables et des effets calculés pour le modèle multilinéaire de l'indicateur FEC pour les maisons uniquement. Les performance d'estimation moyennes sont également données. Source : Auteur après calculs sur la base PHEBUS.

	Terme	Effet estimé	Incertitude	P-valeur
	(Intercept)	4,00	0,92	0%
HS_TYPEHEAT_C (Réf: Electrique individuel)	Fioul individuel	0,16	0,03	0%
	Gaz individuel	0,17	0,03	0%
	Pompe à chaleur	-0,03	0,03	31%
	Bois	0,21	0,02	0%
	Autre	-0,07	0,03	1%
HS_SURF_N		0,55	0,05	0%
DPE_N		0,22	0,03	0%
TC_US5_C		0,02	0,01	0%
WL_EQ1_N		0,01	0,00	0%
HS_AGEBOILER_C (Réf: <1991)	1991-2001	-0,07	0,02	0%
	2001-2012	-0,02	0,02	27%
	NO_INDBOILER	-0,01	0,01	25%
TC_US1_N		0,02	0,00	0%
F_EQ1_N		0,01	0,01	1%
LI_EQ1_C (Réf : Halogènes: Non)	Halogènes: oui	0,04	0,01	0%
HS_STATUS_C (Réf: Autres)	Propriétaire	-0,08	0,03	2%
	Locataire - secteur privé	-0,05	0,04	23%
	Locataire - secteur public	-0,13	0,05	0%
DJU_N		0,00	0,00	0%
HH_DATE_N		0,00	0,00	1%
HH_COMPO_C (Réf: Couple sans enfant)	Couple avec enfants	0,01	0,01	55%
	Famille monoparentale	-0,05	0,03	3%
	Plusieurs personnes	-0,05	0,04	29%
	Personne seule	-0,04	0,02	2%
TC_REG4_C	Non	0,03	0,01	1%
	Parfois	-0,03	0,02	28%
HS_DATECONSTR_C (Réf:<1919)	1919-1945	-0,04	0,02	4%
	1946-1970	0,03	0,02	11%
	1971-1990	0,00	0,02	94%
	1991-2005	0,01	0,02	47%
	>2005	-0,02	0,01	10%
HS_TYPE_C (Réf: Maison multifamiliale)	Maison individuelle	-0,03	0,03	33%
	Maison mitoyenne	-0,06	0,03	8%
HY_EQ1_N		0,01	0,01	5%
TC_EQ2_C (Réf: Régulation centralisée et non prog.)	Régulation centralisée et programmable	0,00	0,02	82%
	Régulation décentralisée	0,03	0,02	17%
	Pas d'éq. De régulation	-0,01	0,02	58%
HS_INDIVMETER_C	Compteur individuel : oui	0,08	0,04	9%
OCC2_N		0,00	0,00	10%

R^2	RMSE	MAE	MAPE
38% [5%]	9417 kWh [750 kWh]	6482 kWh [422 kWh]	31% [2%]

L'analyse comparée des deux modèles permet de faire plusieurs remarques. En premier lieu, les performances d'estimation sont très différentes : si pour les appartements l'erreur quadratique d'estimation est de 6,5 MWh et le R^2 est de 59% en moyenne, pour les maisons, l'erreur est bien plus grande avec un RMSE de 9,5 MWh et un R^2 moyen de 38%. La qualité de l'estimation semble bien meilleure pour les appartements que pour les maisons individuelles. Aussi, l'analyse des coefficients permet d'observer que l'effet de la variable surface est plus important pour les appartements que pour les maisons individuelles (0,88 contre 0,55). De même la qualité thermique des logements semble influencer la consommation en énergie finale des appartements de manière beaucoup plus importante que pour les maisons (0,47 versus 0,22). Enfin, le modèle de consommation pour les maisons semble

dépendre de manière significative du nombre de personnes et de la composition du ménage occupant, contrairement au modèle restreint aux appartements.

Un calcul similaire a été réalisé sur deux échantillons restreints aux logements chauffés aux gaz et aux logements chauffés à l'électricité. Les tableaux de résultats ne sont pas repris ici mais l'analyse montre que des observations similaires peuvent être faites : les variables sélectionnées ne sont pas les mêmes (le modèle restreint aux logements chauffés au gaz est ainsi très lié aux variables de régulation du chauffage, tandis que le second ne l'est que très peu). Aussi, les coefficients associés aux effets des variables sont différents, notamment pour la surface et le DPE.

2.1.4 Conclusion sur les modèles linéaires

L'analyse des résultats permet de voir un écart dans les sélections des variables et les effets moyens calculés entre les modèles de l'indicateur FEC sur l'échantillon total et sur des sous-échantillons (appartements/maisons et chauffage électrique/chauffage gaz). Aussi, on observe que lorsque que l'on réduit la diversité de l'échantillon selon ces deux variables (type de logement et mode de chauffage), on voit apparaitre un nombre plus important de variables de comportement pour expliquer les consommations d'énergie. Le changement d'échelle opéré met en avant des variables d'équipement, d'occupation et de régulation thermique et modifie les effets moyens calculés. L'effet de la surface moyen pour le parc estimé était de 0,56 et est de 0,88 pour les appartements seuls et de 0,55 pour les maisons individuelles. L'effet du DPE est de 0,23 pour le parc, de 0,47 pour les appartements et de 0,22 pour les maisons. De manière intéressante, certaines variables voient leur effet moyen inchangé par le changement d'échelle comme la période de chauffe (TC_US5_C avec effet moyen de 0,02), le niveau d'équipement (F_EQ1_N, HY_EQ1_N).

La modélisation linéaire de la CED a été explorée dans cette partie afin de mieux connaitre les liens entre les variables et 3 indicateurs de la CED. L'étude a souligné deux résultats qui sont (1) le caractère déterminants de variables liées aux contextes résidentiels et (2) la dépendance des effets et des variables à l'échantillon et l'échelle spatiale considérée. Le lecteur notera les écarts significatifs des performances d'estimation entre les modèles construits : 42% de la variance de FEC est expliquée par des variables essentiellement liées aux contextes résidentiels, tandis que 59% de la variance (resp. 38%) était expliqué dans les cas des appartements (resp. maisons) par des variables décrivant le système de chauffage, la surface, et les pratiques des ménages.

Ces résultats amènent alors deux interrogations :

- Si les variables explicatives optimales dépendent d'un sous-échantillon, quelle serait la performance d'un estimateur qui serait « libre » dans la construction des échantillons ? De manière plus quantitative : quelle serait l'erreur (en termes de R^2, de RMSE) d'un tel modèle ? Quel serait le

poids joué par les variables explicatives sélectionnées et en particulier celle de comportements dans un tel modèle ?

- La seconde interrogation est celle du calcul du sous-échantillon « optimal ». Dans cette partie nous avons montré que le fait de restreindre l'échelle permettait de sélectionner des variables explicatives et de calculer des effets moyens plus moins différents par rapport à l'échantillon total. En mettant en relation cela avec la littérature sur les pratiques domestiques et les travaux développés dans le chapitre précédent, nous pouvons nous demander quelle seraient les variables et les effets calculés sur des échantillons homogènes en termes de situations d'habitation ?

Nous traitons chronologiquement ces deux interrogations dans la suite de ce chapitre.

2.2 Identification de la performance d'estimation maximale sur les données PHEBUS : modélisation non linéaire par apprentissage

Avant de construire un modèle qui articulerait ménages, logements, pratiques et consommations d'énergie, il nous semble important de connaître quel est le « pouvoir explicatif » contenu dans les données à disposition. Les questions liées sont : a-t-on « suffisamment » de variables pour décrire les phénomènes qui génèrent des consommations d'énergie ? Les variables dont nous disposons décrivent elles « suffisamment bien » ces processus ? A titre d'exemple, on peut dire que les enquêtes dont nous disposons nous renseignent sur des comportements moyens de chauffage et de régulation, et nous donnent des caractéristiques physiques élémentaires du logement telles que la surface, le nombre de pièce. Néanmoins en dépit de cela, plusieurs paramètres peuvent générer de grandes variations dans la CED et les pratiques des ménages, comme la ventilation du logement qui peut être passive (l'air est renouvelé par les ouvrants et les fissures) ou active (une ventilation mécanique contrôlée assure le renouvellement). Il est possible que les données à disposition permettent d'expliquer 100% de la variance des indicateurs de CED (soit aussi une erreur quadratique nulle), mais il est également probable que le « bruit » de mesure (erreur dans la réponse, sa retranscription, pas de réponse) et le manque de variables nuisent à la qualité des estimateurs construits. L'idée est alors de lever l'hypothèse de linéarité et de construire un modèle non linéaire qui maximiserait la performance d'estimation.

Durant ce travail de thèse, nous avons conduit un ensemble de manipulations qui ont permis de tester plusieurs stratégies de traitement des caractéristiques (en anglais *feature extraction,* dont l'analyse factorielle, autoencodage de données mixtes) et d'algorithmes (Decision Trees, Random Forest, XGBoost). Par souci de simplicité dans l'exposé, nous ne présentons ici que le modèle ayant présenté la meilleure performance. Cette approche devrait nous permettre de savoir quel pourcentage de variance de la CED est explicable avec les données à disposition.

2.2.1 Intérêt de la méthode XGBoost

Il existe de nombreux modèles d'apprentissage supervisé présentant de bonnes performances d'estimation. Parmi eux nous pouvons citer les modèles basés sur les réseaux de neurones qui fournissent d'excellents estimateurs de la CED comme nous avons pu le voir dans le chapitre 1. Dans ce cas nous avons préféré opter pour un modèle XGBoost (Chen et Guestrin 2016). Nous en détaillons rapidement le principe avant de justifier son usage dans notre cas.

Le modèle XGBoost (contraction de eXtreme Gradient Boosting en anglais) désigne l'implémentation d'un algorithme de « gradient boosting ». Il est basé sur des arbres de décision pour résoudre des problèmes de classification et de régression. Ce modèle a été développé dans les années 2010 et est très reconnu dans la communauté de l'apprentissage pour ses très bonnes performances prédictives. Il y a deux idées fondamentales pour comprendre l'intérêt et le fonctionnement de cet algorithme. La première est le concept de « modèle d'ensemble » : pour produire une estimation d'une variable qualitative quantitative, selon le problème posé, XGBoost se base non pas sur le calcul d'un seul modèle mais sur le calcul de dizaines voire souvent de centaines de « petits » modèles, dont les prédictions sont agrégées (par exemple par un vote). Le terme « petits » intervient ici pour désigner le fait que chacun de ces sous-modèles est un modèle spécialisé car entraîné sur un sous-échantillon de la base de données. Dans notre cas chacun des sous-modèles est un arbre de décision, entraîné sur un nombre limité de lignes et de colonnes de la base de données (par exemple 30% des lignes et 30% des variables).

La seconde idée est celle du « boosting » qui consiste à construire chacun des sous-modèles à partir des erreurs de prédiction du sous-modèle qui le précède (Figure 40), ce qui a pour effet de donner plus de poids aux valeurs difficiles à prédire lors de la phase d'apprentissage. L'objectif poursuivi est d'améliorer la précision du modèle final.

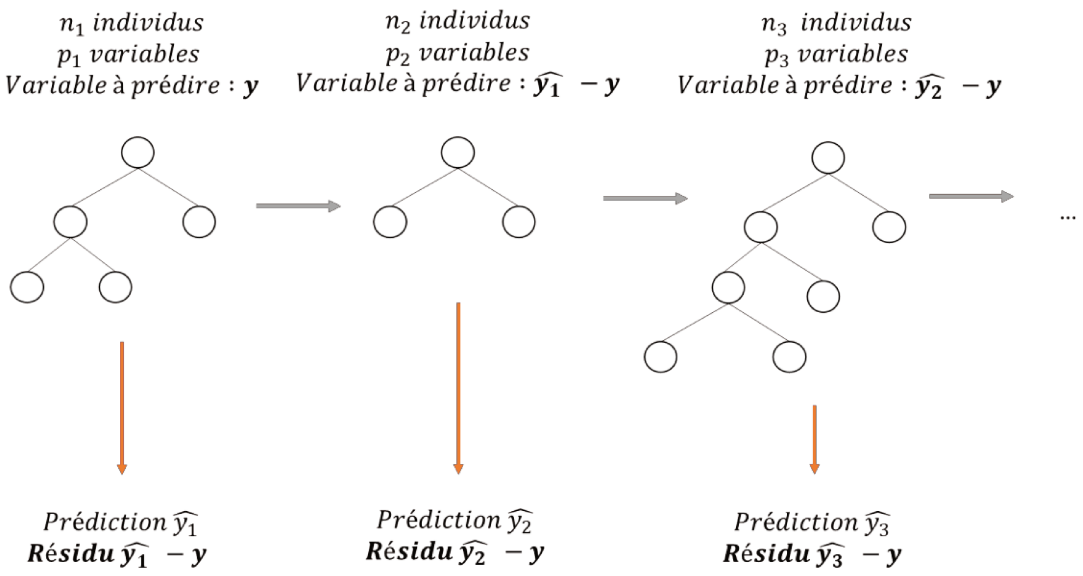


Figure 40 : Principe général de l'algorithme XGBoost, basé sur des arbres de décision. Source : Auteur, adapté de (Chen et al. 2016).

Le choix de ce modèle est lié d'un part à sa bonne performance prédictive ainsi qu'à une meilleure interprétabilité des résultats comparé aux modèles basés sur les réseaux de neurones. Il est par exemple possible d'observer la structure des chacun des arbres. Il existe par ailleurs des métriques dans la littérature permettant de mesurer le poids des variables dans la construction des estimations.

Ses faiblesses sont le fait qu'il existe un risque important d'*over-fitting*. Ce phénomène désigne le fait qu'un modèle a appris une structure mathématique trop rigide pour pouvoir fournir une estimation précise sur de nouvelles données. On observe dans ce cas que les métriques mesurant l'erreur du modèle sont excellentes sur les données d'entraînement mais médiocres sur le *set* de test. La seconde faiblesse est le nombre important de paramètres du modèle. Nous citerons :

- Nt : Le nombre d'arbres, c'est-à-dire le nombre d'itérations de *boosting*.
- Dt : La profondeur, c'est-à-dire la taille maximale des arbres.
- Pc : Le pourcentage de colonnes de la base de données sélectionnées aléatoirement pour l'entrainement de chaque sous-modèle.
- Pl : Le pourcentage de lignes retenu aléatoirement pour entrainer chaque arbre.
- Eta : Il s'agit d'un paramètre contrôlant la vitesse d'apprentissage du modèle.

Dans notre étude, nous recherchons un paramétrage optimal en testant un grand nombre de combinaison de ces paramètres et nous retenons le paramétrage pour lequel les métriques de performance sont optimales. Nous utilisons dans cette étude la librairie *xgboost* (Chen et *al.*, 2023).

2.2.2 Entrainement du modèle XGBoost et analyse des résultats

Le modèle optimal retenu est paramétré de la manière suivante : Nt= 100 arbres ; Dt = 3 ; Pc = 50% ; eta = 0,1 ; Pl = 90%. Les performances d'estimation sont calculées par moyenne après entrainement et évaluation sur 50 échantillons tirés aléatoirement. On donne ici les performances sur l'ensemble de test après avoir vérifié que ces performances n'étaient pas trop différentes des valeurs observées sur le *set* d'entraînement.

R^2	RMSE	MAE
45.2 % (4.5%)	9149 kWh (535 kWh)	6152 kWh (274 kWh)

On observe que les performances de modèle sont équivalentes voire légèrement supérieures (en termes de variance expliquée) à celles du modèle multilinéaire de l'indicateur FEC. Ceci est un argument supplémentaire pour soutenir l'idée que les données à disposition ne contiennent pas les informations suffisantes pour permettre une bonne estimation de la consommation d'énergie FEC. On peut toutefois aussi s'intéresser aux variables sélectionnée par le modèle. La modélisation par XGBoost permet en effet de mesurer « l'importance » des variables dans l'estimation (Figure 41). La surface, le DPE, le type de chauffage y jouent un rôle primordial, avant d'autres variables telles que la période de chauffage (TC_US5_C), la date d'arrivée du ménage dans le logement, le nombre de pièces, l'âge de la PR etc.

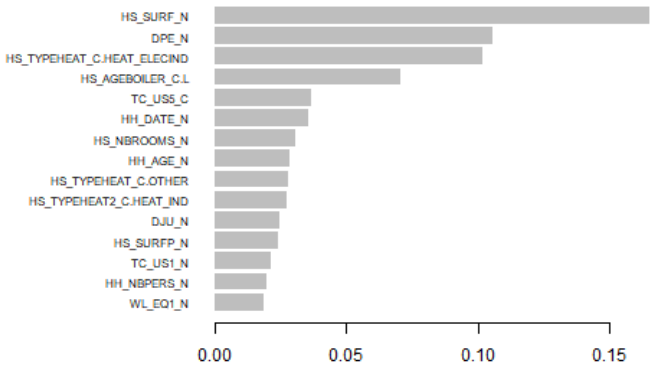


Figure 41 : Mesure de l'importance des variables dans la prédiction des valeurs du modèle XGBoost. Source : Auteur après calculs sur la base PHEBUS.

3. Modélisation par un modèle de mélange de modèle de régression incluant un processus logistique caché

Dans cette partie nous présentons un modèle de la consommation d'énergie domestique. Une première partie permet de rappeler la motivation initiale et de justifier la structure mathématique que nous avons choisie.

3.1 Intérêt de l'approche

La littérature nous montre que la consommation d'énergie dépend (au moins) de la surface, de l'isolation, du type de ménage et de logement, des pratiques domestiques. Les calculs de régression linéaire et du modèle XGBoost nous permettent de confirmer cela mais aussi de rappeler que les coefficients de régression identifiés semblent dépendre de la composition de la base de données considérée, notamment en termes de type de logement.

Cette observation nous a amené à considérer qu'il fallait au moins différencier les types de logement pour construire un modèle de consommation d'énergie. D'autres variables semblaient toutefois interagir avec celle-ci : elle est en effet liée au revenu, au type de ménage, à son âge, et au contexte résidentiel. Cette observation nous a conduit à considérer qu'il fallait segmenter plutôt l'espace des situations d'habitation pour préserver notre capacité à interpréter les situations d'habitation identifiées par le calcul de la segmentation. A cette étape du travail, la question est méthodologique : comment parvenir à une segmentation « optimale » des situations d'habitation pour construire un modèle de régression de la consommation d'énergie ? Comment calculer cette régression ?

Une première approche consisterait comme nous l'avons fait dans la partie précédente à procéder en deux temps : d'abord calculer une partition des situations d'habitation puis modéliser la consommation d'énergie sur ces sous-ensembles de données, par exemple avec une régression multilinéaire. L'avantage de cette approche est qu'elle permet de mettre en évidence la sensibilité des coefficients et des effets « toutes choses égales par ailleurs » pour l'ensemble des situations d'habitation considérées (Figure 42). En revanche, cette méthode suppose de définir deux critères. Un premier qui permette de calculer une partition des situations d'habitation et un second pour calculer la régression sur chacun des segments. Aussi, un grand nombre de classes diminuerait la significativité statistique des coefficients calculés. Cette approche permet certainement par itération d'identifier un nombre restreint de classes et des coefficients significatifs. Toutefois, nous avons jugé intéressant d'étudier une méthodologie alternative qui propose une formulation mathématiquement plus parcimonieuse.

Dans cette alternative nous proposons de réaliser simultanément le partitionnement et la régression. L'avantage de ce type d'approche est qu'il limite l'*apriori* du modélisateur à la sélection des variables, à la structure du modèle de régression et au choix du critère de performance. La Figure 42 montre deux voies existantes permettant de calculer un modèle de régression sur des segments homogènes de situations d'habitation. Nous développons dans la suite de ce chapitre une approche basée sur le modèle RHLP (acronyme de l'anglais « Regression model with a Hidden Logistic Process »).

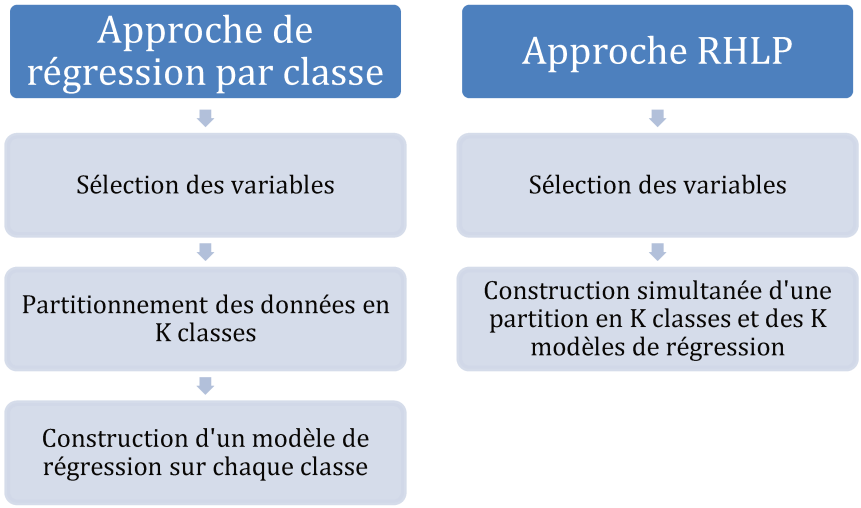


Figure 42 : Comparaison des approches classiques et "intégrées" de régression sur des sous-ensembles de données. Source : Auteur.

Dans cette formulation, la modélisation des consommations repose sur l'identification de classes de « situations d'habitation » et l'estimation de coefficients de régression.

3.2 Description de l'algorithme RHLP

L'algorithme RHLP est un modèle probabiliste qui a été créé initialement pour segmenter des séries temporelles. L'idée de ce modèle est de valoriser le lien entre un ou plusieurs variables expliquées (notées $\boldsymbol{Y}$) et des variables explicatives (notées $\boldsymbol{X}$), comme le temps, pour identifier des changements de régime et les caractériser à l'aide de paramètres de fonctionnement (calculés par le modèle). La Figure 43 est tirée de la documentation publiée par Faicel Chamroukhi à propos d'une version de l'algorithme en langage R. L'exemple permet d'observer un signal temporel $\boldsymbol{Y}$ évoluer (en noir sur la figure). L'utilisation de l'algorithme RHLP permet alors d'identifier cinq phases où les évolutions temporelles de $\boldsymbol{Y}$ sont statistiquement très différentes. Le modèle permet alors aussi de calculer des modèles (en couleur sur la figure) sur chacune des phases ce qui permet de caractériser les phases à partir de l'analyse des coefficients (croissance, stabilité etc.).

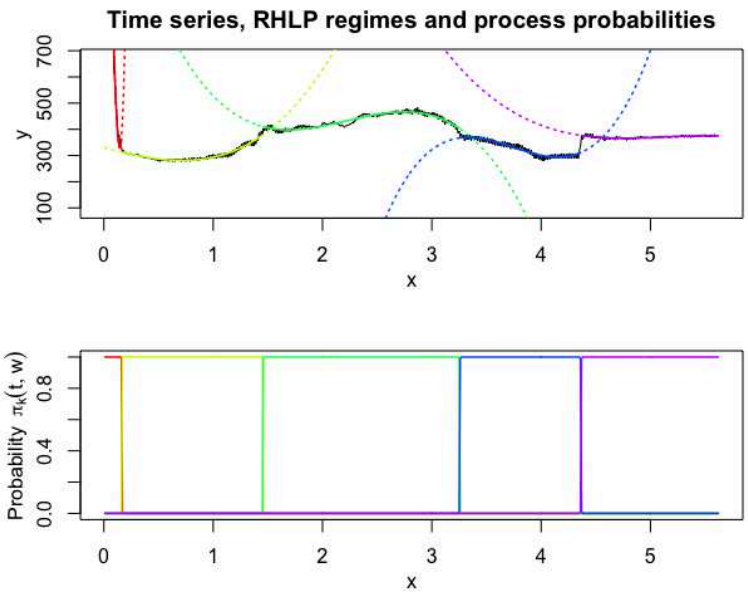


Figure 43 : Illustration de l'utilisation du modèle RHLP pour l'identification de régimes à partir de séries temporelles. Dans cet exemple cinq régimes sont identifiés. Le graphique du haut montre les données mesurées ***y*** *en fonction du temps* ***x****. En couleur, on trace les modèles de régression (polynomiaux et fonction de* ***x)*** *sur chacun des cinq segments. Les traits pointillés correspondent aux valeurs prises par les modèles en dehors des segments sur lesquels ils sont définis initialement. Sur le graphe du bas, on peut lire la probabilité pour chaque valeur de* ***x*** *d'être dans le cluster* i ($i\epsilon[1,5]$). *L'illustration et l'exemple sont tirés de la documentation du package* ***samurai*** *sur le site CRAN[30]. Source : Documentation du package samurai tirée du site CRAN, à partir de (Chamroukhi, 2009).*

Ce premier exemple est décrit ici de manière qualitative mais il permet de se rendre compte de l'apport méthodologique de cet algorithme. Il permet de construire simultanément un partitionnement (ici une détection de changement de régime) et un modèle hiérarchique parcimonieux de la variable expliquée (ici utiles ici par exemple pour étudier les caractéristiques physiques associées à chacun des régimes).

Ce modèle correspond à ce que nous voulions faire sur les données de consommation d'énergie et en utilisant les variables caractérisant les situations d'habitation comme variable explicatives (*X*). Il nous a paru intéressant aussi d'étudier l'usage de ce type d'algorithme sur des données mixtes d'enquête. Dans notre approche, il s'agit alors non plus de segmenter des séries temporelles mais de partitionner les situations d'habitation et de construire des modèles de régression sur chacune des classes. Dans un premier temps nous faisons le choix de construire sur chaque segment un modèle de régression linéaire.

On parlera dans la suite plutôt d'une variante de l'algorithme RHLP qui est l'algorithme MRHLP (pour « Multivariate Regression model with a Hidden Logistic Process »). Ce dernier diffère du modèle RHLP dans la mesure où il calcule non plus une régression d'un seule variable expliquée mais de plusieurs variables expliquées simultanément.

[30] Le package samurai est une librairie en langage R permettant d'implémenter une version du modèle RHLP. Accès au site via l'URL : https://cran.r-project.org/web/packages/samurais/readme/README.html

On considère quelques éléments de notation :

- $\boldsymbol{y_i} = (y_{1i}, y_{3i}, y_{3i})$ le vecteur composé des variables expliquées (dépendantes). Il peut être de dimension 1 (composé alors d'un seul indicateur de consommation) ou de dimension 3. Dans ce cas il est composé des trois logarithmes décimaux des consommations d'énergie du ménage i : totale (en kWh), par mètre carré (kWh/m²) et par personne (kWh/p).
- $\boldsymbol{x_{1,i}}$ le vecteur de dimension p_1 composé des variables de partitionnement décrivant le ménage et le logement i.
- $\boldsymbol{x_{2,i}}$ le vecteur de dimension p_2 composé des variables de régression décrivant le ménage et le logement i.
- $\boldsymbol{\beta}_{z_i}^{(j)}$ le vecteur de dimension p_2 composé des coefficients de régression pour le paramètre j du vecteur $\boldsymbol{y_i}$ du modèle k et décrivant le ménage et le logement i.
- $\boldsymbol{\sigma}_{z_i}^{(k)}$ le vecteur des variances.
- $\boldsymbol{z} = (z_i)_{i\epsilon[1,n]}$ la variable (cachée) décrivant l'appartenance du ménage i à l'un des classes.

On peut alors écrire pour chaque segment k, et pour chacune des 3 variables expliquées un modèle de régression pour les 3 consommations :

$$\forall j \in [1,3] \qquad y_k^{(j)} = \boldsymbol{\beta}_k^{(j)}.\boldsymbol{x_2} + \sigma_k^{(j)}\epsilon$$

Ecrit sous forme matricielle, cela donne :

$$\boldsymbol{y}_k = \boldsymbol{\beta}_k.\boldsymbol{x_2} + \boldsymbol{\Sigma}_k$$

Où $\boldsymbol{\Sigma}_k$ est la matrice de covariance du groupe k. L'appartenance d'un ménage i à un segment k est codée à travers la variable z_i, qui est une variable dite « cachée » c'est-à-dire que sa valeur n'est pas observable mais qu'elle est calculée lors de la phase d'entraînement du modèle.

Calcul de la partition en K classes

L'entraînement du modèle repose sur un critère probabiliste. On écrit la probabilité d'observer les données de consommation d'un ménage $\boldsymbol{y}_i$ et son appartenance au segment z_i conditionnellement au modèle $\mathcal{M}$ et ses paramètres $(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k, \boldsymbol{w}_k)_{k\in[1,K]}$:

$$p(\boldsymbol{y}_i, z_i \,|\mathcal{M}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x_{1,i}}, \mathbf{w}_k)\, \mathcal{N}(\boldsymbol{y}_{ik}|\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)$$

Où $\pi_k(\boldsymbol{x_{1,i}}, \mathbf{w}_k)$ est la probabilité pour le ménage i d'appartenir au segment k. $\mathbf{w}_k$ désigne le vecteur des poids (propres au cluster k) permettant de calculer π_k . π_k peut être considéré comme le « degré d'appartenance » du ménage au segment k.

$$\pi_k(\boldsymbol{x_{1,i}}, \mathbf{w}_k) = \frac{exp(\sum \boldsymbol{x_{1,i}}, \mathbf{w}_k)}{\mathbf{1} + exp(\sum \boldsymbol{x_{1,i}}, \mathbf{w}_k)}$$

On peut alors définir la vraisemblance $L(\mathcal{M})$ du modèle conditionnellement au modèle $\mathcal{M}$ comme la probabilité d'observer l'ensemble des données de consommation des ménages, conditionnellement au paramétrage du modèle.

$$L(\mathcal{M}) = p(\boldsymbol{y}, \boldsymbol{z} \,|\mathcal{M}) = \prod_{i=1}^{N} p(\boldsymbol{y}_i, z_i \,|\mathcal{M})$$

Choix du modèle MRHLP

On a ici défini la **vraisemblance** d'un modèle qui permet d'évaluer et de comparer quantitativement la capacité de plusieurs modèles à restituer la structure des données d'un ensemble de données. On garde cependant bien en tête qu'il est possible à ce stade de choisir de nombreux paramètres : le nombre de classes K, les variables $\mathbf{X_1}$ servant à construire un partitionnement et les variables servant à calculer une régression $\mathbf{X_2}$. Parmi tous les modèles réalisant ces différentes combinaisons de paramétrage, on pourrait choisir le modèle qui maximiserait ce critère de vraisemblance. En pratique, les modèles qui ont plus de paramètres ont cependant plus de facilité à maximiser ce critère : ils sont donc avantagés par construction. C'est pourquoi les modélisateurs maximisent plutôt $p(\mathcal{M}|\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{y}, \boldsymbol{z} \,|\mathcal{M}) \frac{p(\mathcal{M}, \boldsymbol{y}, \boldsymbol{z})}{p(\mathcal{M})}$.

Dans le chapitre précédent, nous avons parlé du critère ICL qui est une extension des critères AIC (Critère d'information d'Akaike – en anglais Akaike Information Criterion) et BIC (Critère d'Information Bayésien – en anglais Bayesian Information Criterion) (Bishop 2006) qui sont plus courants dans la littérature. Ces critères sont des expressions asymptotiques permettant de maximiser l'expression ci-dessus. Ils diffèrent dans les hypothèses prises pour calculer les expressions analytiques. Ces critères servent à pénaliser les modèles qui ont « trop » de paramètres en modélisant la perte d'information. Les calculs amenant à leur définition ne sont pas rappelés ici mais le lecteur pourra en apprendre plus en se référant par exemple à l'article de (Kuha 2004). La différence entre les critères est que BIC prend en compte la taille de l'échantillon (N) pour calculer cette pénalité.

$$AIC = -2\ln(L) + 2k$$

$$BIC = -2\ln(L) + k.\ln(N)$$

L'usage et la comparaison des critères dans notre processus de construction du modèle a montré que le critère AIC permettait de construire des modèles plus intéressants car fournissant à la fois des partitions intéressantes et de bonnes performances d'estimation. Ces critères probabilistes peuvent être couplés aux critères classiques de la littérature sur les modèles de consommation d'énergie et utilisés précédemment (R^2, RMSE, MAE, MAPE).

Calcul de la régression

La particularité de ce modèle est qu'il contient k sous-modèles pondérés par des poids π_k. Le calcul des valeurs estimées des valeurs des 3 variables expliquées peut se faire de deux manières, en faisant des hypothèses sur les valeurs des poids à choisir.

- En considérant que la valeur estimée pour un individu i vient uniquement de la contribution du sous-modèle k auquel il est le plus lié :

$$\hat{y}_i^{(j)} = \boldsymbol{\beta}_{z_i}^{(j)} . \boldsymbol{x_2}$$

- Il est cependant possible de prendre en compte le fait que le modèle peut être lié à deux groupes. Dans cette perspective, on pondère les contributions de chacun de ces sous-modèles :

$$\hat{y}_i^{(j)} = \sum_{k=1}^{K} \pi_k(\boldsymbol{x_{1,i}}, \mathbf{w}_k) . \boldsymbol{\beta}_k^{(j)} . \boldsymbol{x_{2,i}}$$

Par expérience, la seconde modélisation fournit de meilleures estimations (en termes de précision) : elle tire parti de la souplesse permise par la pondération des sous-modèles calculés.

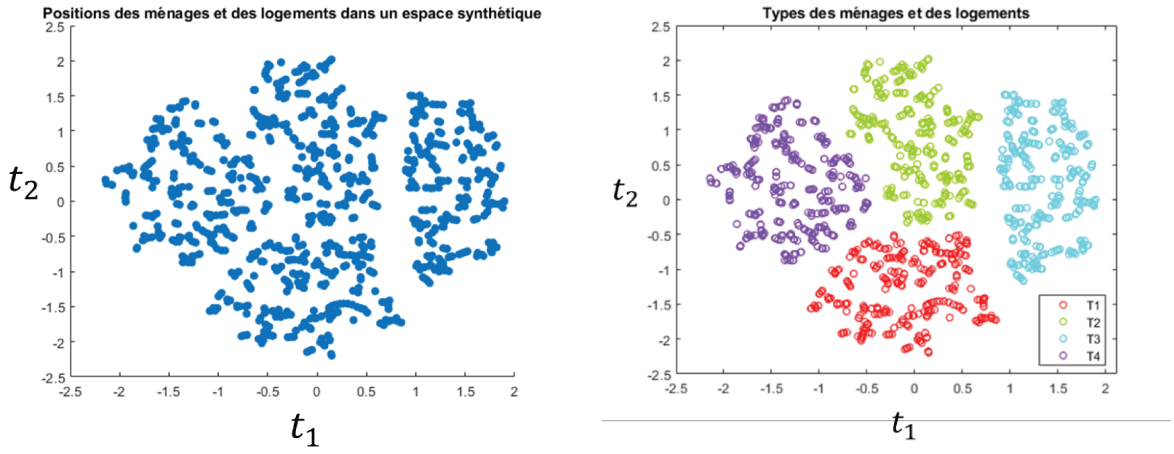
Evaluation du modèle

L'évaluation du modèle est effectuée à l'aide de critère quantitatifs qui permettent de caractériser la qualité de l'estimation du modèle. Les indicateurs R^2 et RMSE seront utilisés.

Cependant, en gardant à l'esprit notre volonté de construire un modèle explicatif nous évaluons aussi qualitativement (1) les clusters de situations d'habitation, (2) les pratiques des ménages rattachés à chacun des clusters, (3) les coefficients de régressions calculés et (4) les niveaux de consommation en énergie finale pour les trois indicateurs. Le modèle est jugé acceptable s'il est possible de donner du sens à ces éléments. Il est jugé meilleur qu'un autre modèle si ses performances d'estimation sont meilleures et que l'interprétation donne des résultats satisfaisants.

3.3 Un exemple introductif

Considérons un exemple fictif où il existerait 4 groupes (T1, T2, T3, T4) au sein d'un set de données. Chacun des groupes contient 300 individus. Supposons qu'il soit possible de les représenter dans un espace à deux dimensions $(t1, t2)$. On a donc une matrice de dimension 1200x2 qui permet de positionner les individus dans un espace. Ces 4 groupes sont « presque » identifiables sur la Figure 44 (a), mais leur identification à leur type « vrai » n'est pas évidente (figure b).



(a) (b)

Figure 44 : Représentation d'un ensemble de données où 4 classes sont présentes. L'espace synthétique comprend deux dimensions : t1 (abscisse) et t2 (ordonnée). Source : Auteur

Un algorithme de partitionnement simple comme la classification ascendante hiérarchique ou l'algorithme K-means pourrait être utile pour identifier les 4 classes (T1, T2, T3, T4). Nous proposons d'adopter une approche légèrement différente dans le sens où pour « aider » l'algorithme à identifier ces types nous allons lui donner une information supplémentaire sur les types que nous recherchons : nous savons que au sein de chacune des classes il existe un lien différent entre 6 variables descriptives que nous noterons $(X_1, X_2, X_3, X_4, X_5, X_6)$ et une variable expliquée Y. Cette information supplémentaire va aider l'algorithme à discriminer les groupes. On donne sur la Figure 45 une représentation dans l'espace (t_1, t_2) de la distribution de cette variable. On observe qu'elle prend des valeurs moyennes différentes selon les segments.

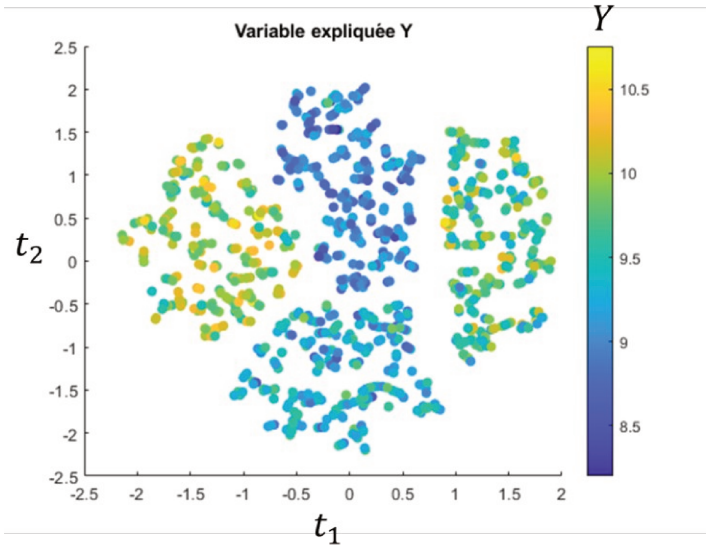


Figure 45 : Position des données dans l'espace synthétique (t1, t2). Un code couleur est utilisée pour représenter la distribution de la variable dépendante. Source : Auteur.

Notre objectif est alors d'identifier ces groupes conditionnellement au fait que les individus vérifient une relation commune à chacun des groupes avec une variable expliquée Y. Comme variables de partitionnement, nous utilisons la matrice $\boldsymbol{X_1} = \left[\mathbf{1}, \boldsymbol{t_1}, \boldsymbol{t_2}, \boldsymbol{t_1^2}, \boldsymbol{t_2^2}\right]$. Ce choix de modélisation permet de construire des classes sphériques dans l'espace $(\boldsymbol{t_1}, \boldsymbol{t_2})$. Comme variable de régression, nous utilisons la matrice $\boldsymbol{X_2} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}]$.

Autrement dit, en se rappelant que la variable z_i indique l'appartenance de l'individu i à un groupe :

$$\exists(\beta_{z_i}, \sigma_{z_i})_{z_i \in [1,4]} \; tel \; que \; \forall i \quad y_i = \beta_{z_i}.\boldsymbol{x_{2,i}} + \sigma_{z_i}\epsilon_i$$

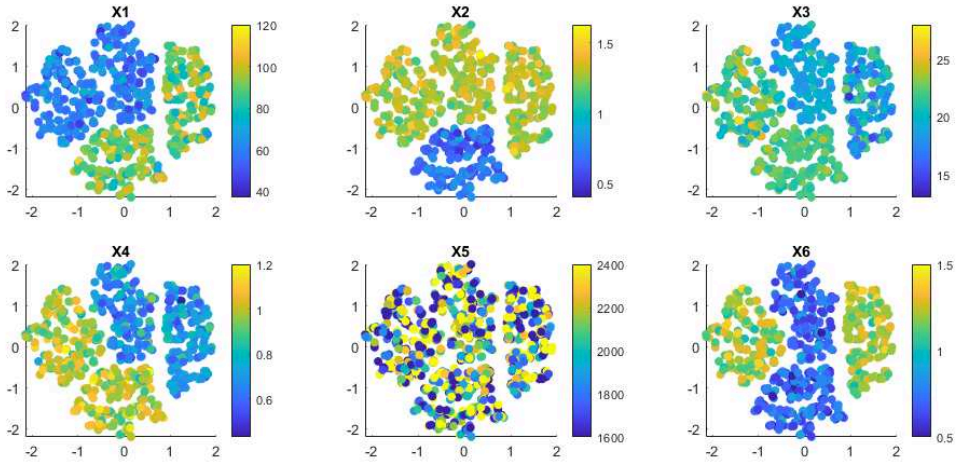


Figure 46 : Valeurs prises par les valeurs explicatives. Chaque figure permet de visualiser la distribution (en couleur) des valeurs prises par les variables X_i . Source : Auteur.

L'entraînement du modèle MRHLP est effectué 20 fois. Nous recherchons 4 classes mais nous effectuons ce calcul pour plusieurs valeurs de K afin d'observer l'évolution des indices AIC et BIC. On trace les évolutions des indicateurs AIC et BIC sur la Figure 47. On observe que les indices décroissent avec K. Nous pourrions sélectionner ainsi K=6 ou 7. Néanmoins, l'observation des coefficients β montre que les classes supplémentaires créées sont très proches des autres : l'ajout de classes n'apporte pas d'information. On retient K=4 pour la suite de l'analyse.

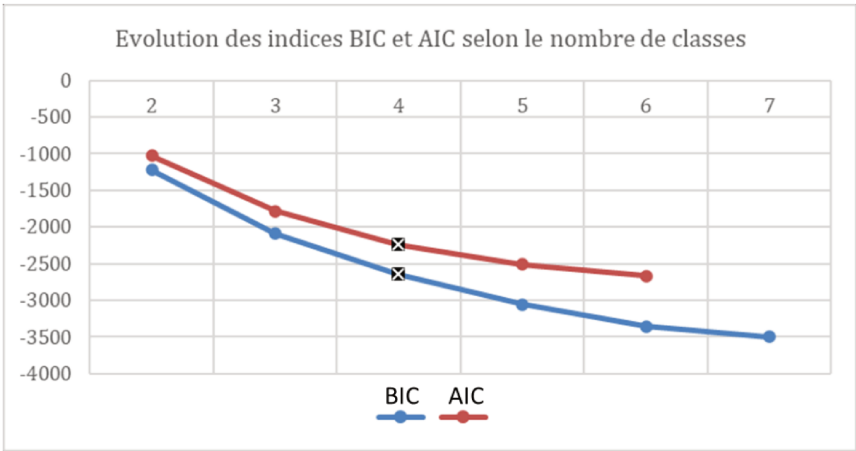


Figure 47 : Evolution des critères AIC et BIC en fonction du nombre de clusters K. Source : Auteur.

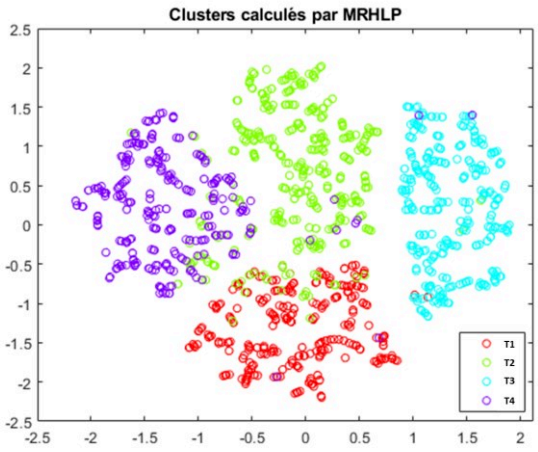
Pour évaluer le modèle on procède en 3 temps : l'analyse du partitionnement, l'analyse de la performance de l'estimateur, l'analyse des coefficients de régression.

Analyse du partitionnement

L'appartenance à la classe k est calculé à l'aide des facteurs contenus dans le vecteur $\boldsymbol{\pi_k}(t_1, t_2)$, qui sont calculés à l'aide de modèles logistiques. Les paramètres de ces derniers (un modèle pour chaque groupe) sont calculés lors de l'entrainement. Ainsi pour chaque point i, on dispose de 4 valeurs $\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4}$ traduisant le degré d'appartenance à chacun des types calculés. On décide que la classe calculée pour le point i est celle qui dont la probabilité *a posteriori* π_i est maximale :

$$z_i = \underset{z_k}{\operatorname{argmax}}\, \pi_k\,(i)$$

On croise alors le partitionnement calculé avec le partitionnement attendu. On utilise pour cela une matrice de confusion qui permet d'analyser les correspondances entre les deux partitions (Figure 48). On observe alors que le modèle offre une performance convenable puisque seuls 6,8% des éléments ne sont pas regroupés comme dans la partition originale. Mathématiquement, on peut calculer la proportion de ces individus « mal classés » à l'aide de l'indice de Rand qui vaut ici 93%, ce qui est un très bon score. On remarque que ce sont essentiellement les classes T1 et T4 qui sont plus difficilement discriminées, et confondues au sein de la classe T2.



		Classes calculées			
		T1	T2	T3	T4
Classes « vraies »	T1	272	26	0	2
	T2	0	294	0	6
	T3	2	2	294	2
	T4	0	42	0	256

Figure 48 : Analyse du partitionnement calculé par MRHLP. A gauche, les ménages sont représentés dans l'espace synthétique et colorés selon la classe attribué. A droite, la matrice de confusion permet d'observer les écarts entre les types vrais et les types calculés. Source : Auteur.

Analyse de la performance d'estimation

Le modèle calculé permet également de calculer des valeurs estimées de la grandeur Y, de la manière suivante :

$$\widehat{\boldsymbol{Y}} = \sum_{k=1}^{4} \pi_k(t_1, t_2) \,.\, \boldsymbol{\beta}_k . \boldsymbol{X}$$

Où $\widehat{\boldsymbol{Y}}$ est le vecteur des valeurs estimées de $\boldsymbol{Y}$ (de dimension 1200x1), $\boldsymbol{\pi_k}$ le vecteur contenant les poids du type k (dimension 1200 x 1), $\boldsymbol{\beta}_k$ la matrice des coefficients de régression de la classe k (dimension 1 x 7) et $\boldsymbol{X}$ la matrice des variables de régression (dimension 1200 x 7)[31]. On calcule plusieurs indicateurs de performance : le R^2 vaut 98,4% sur l'ensemble de test et l'erreur quadratique normalisée (NRMSE) est de 12,4% ce qui dénote un très bon estimateur. Sur la Figure 49, on peut observer une faible dispersion des valeurs prédites par rapport aux valeurs à prédire (à gauche). Les résidus (ie. les écarts entre les valeurs prédite et les valeurs réelles) sont représentés sur la figure de droite et colorés en fonction de la classe et du modèle associé qui réalise l'estimation pour chacun des points. Par exemple, les points de couleur violette sont les résidus calculés pour les points appartenant à la classe T4. A titre de comparaison, nous avons réalisé un modèle de régression multilinéaire et le modèle obtient un R^2 de 94% (également sur l'ensemble de test). Ces performances sont peu dépendantes des échantillons d'entraiment et de test.

[31] X contient 7 colonnes et non 6 car la première colonne contient uniquement des 1.

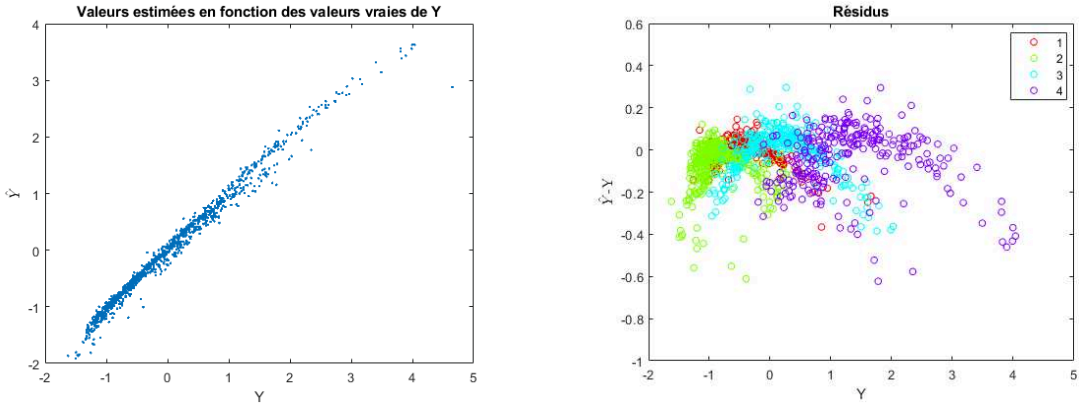


Figure 49 : Tracé des valeurs estimés et des résidus du modèle MRHLP à 4 classes. Source : Auteur

Analyse des coefficients

La comparaison des coefficients calculés par le modèle RHLP sur chacune des classes et des coefficients calculé par un modèle de régression multilinéaire MLR est intéressante. Elle permet d'observer la sensibilité des coefficients du modèle MLR à l'échantillon. Le Tableau 24 recense ces coefficients. On compare les résultats du modèle MLR (colonne de droite) avec les résultats du modèle RHLP. Le modèle MLR identifie une sensibilité faible sur la variable X_1, mais on observe que les différents types ont des sensibilités étalées autour de cet effet moyen (0,19 à comparer avec 0,07 pour la classe 1, 0,33 pour la classe 4). Ensuite, ce tableau de coefficients permet de distinguer des types qui ont une sensibilité moyenne plus élevée à certaines variables de régression. Le type 4 présente une sensibilité plus importante puisque les valeurs des coefficients de régression sont plus importantes. Cette hétérogénéité des coefficients est principalement due à la différence des distributions des X_i dans l'espace synthétique (Figure 46). Cette analyse quantitative permet ensuite d'ouvrir une discussion sur la causalité liant les facteurs X_i avec la variable expliquée Y.

Tableau 24 : Coefficients β issues de l'entraînement du modèle RHLP. Pour chaque type calculé, 7 coefficients sont donnés : le premier donne l'ordonnée à l'origine et les 6 suivants sont associés aux variables explicatives X_i. Source : Auteur.

	Modèle MRHLP - Type				Ensemble
Coefficient	**1**	**2**	**3**	**4**	**(Modèle MLR)**
Origine	0,49	-0,28	0,40	-0,53	0,00
X_1	0,07	0,07	0,09	0,33	0,19
X_2	0,65	0,26	0,51	0,75	0,53
X_3	0,34	0,27	0,51	0,71	0,50
X_4	0,31	0,28	0,55	0,60	0,38
X_5	0,26	0,17	0,35	0,55	0,32
X_6	0,51	0,34	0,33	0,61	0,46

L'analyse de cet exemple nous permet de voir que le fait de disposer d'un espace synthétique permettant de représenter des « individus » (dans notre cas les ménages et les logements, décrits par des contextes résidentiels) permet de décrire à la fois des classes de similarité et l'effet local des variables explicatives.

L'enjeu de la suite de ce chapitre est alors de construire un espace synthétique permettant de représenter les contextes résidentiels puis d'établir un modèle similaire permettant d'estimer la consommation en énergie finale. Les différences (notables) sont principalement relatives à la complexité accrue du modèle : nous avons vu qu'il était difficile de représenter les ménages et les logements dans un espace synthétique (Chapitre 3). Il en résulte que la matrice X_1 comprendra plus que 2 variables. Par ailleurs, ainsi que nous l'avons observé en début de chapitre les consommations totales (FEC), par m² de surface de logement (FEC/m²) et par personne (FEC/p) ont des dynamiques différentes. Il sera alors intéressant de faire une régression non pas uniquement sur FEC mais sur les 3 variables simultanément. Nous discutons de cela dans la partie suivante.

3.4 Méthodologie

3.4.1 Méthodologie suivie pour la construction du modèle global de CED

La construction du modèle MRHLP « global » (pour désigner la modélisation simultanée des 3 indicateurs) repose sur plusieurs étapes présentées dans le schéma suivant (Figure 50). Comme indiqué dans le paragraphe précédent, la modélisation nécessite de représenter les ménages dans un espace synthétique où les situations d'habitation « similaires » sont « proches » : la construction d'un tel espace mathématique et la définition d'une distance associée nécessite des hypothèses de travail qui sont définies au paragraphe 3.5. Ensuite, pour aider à la compréhension du lecteur et pour faciliter le travail nous avons choisi de présenter des modèles MRHLP pour chacun des indicateurs. Les résultats (sélection des variables, qualité des estimateurs) sont présentés au paragraphe 3.6. Le modèle MRHLP globale est présenté quant à lui au paragraphe 3.7. Le détail des opérations mathématiques liées à l'entrainement du modèle (calcul des classes, des coefficients) et à son évaluation est donné au paragraphe 3.4.2.

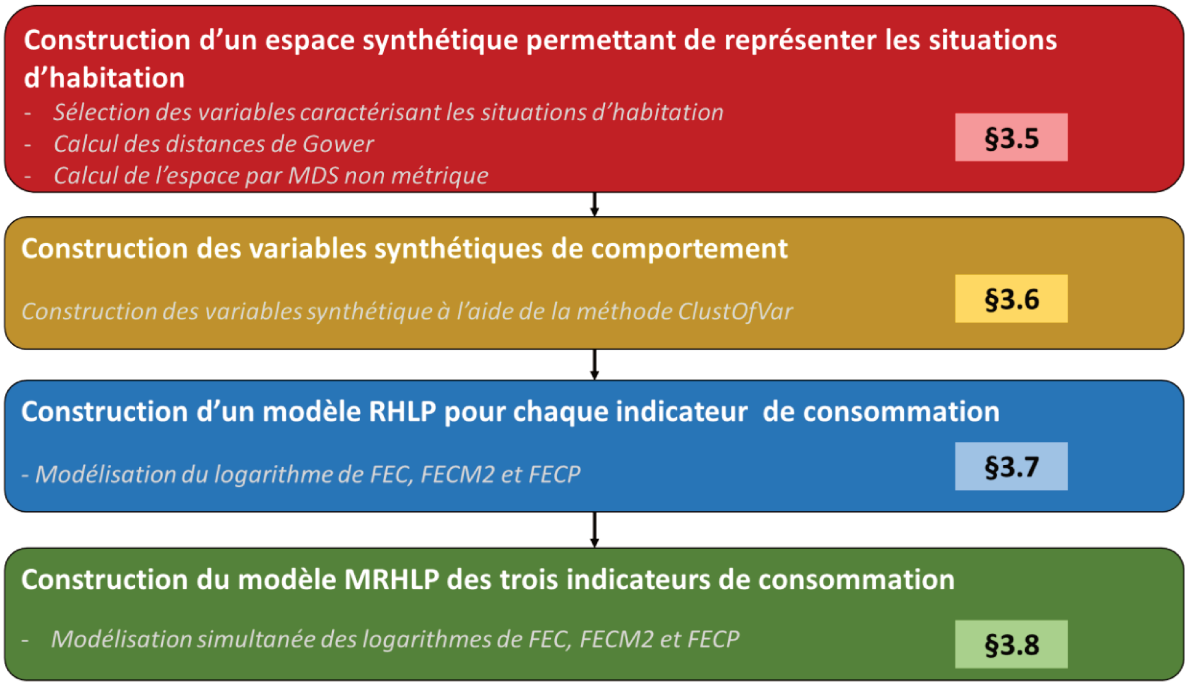


Figure 50 : Méthodologie suivie pour la construction du modèle MRHLP. Source : Auteur.

3.4.2 Méthodologie suivie pour l'entraînement et l'évaluation d'un modèle MRHLP

L'entrainement et l'évaluation d'un modèle MRHLP est basé sur l'article de (Chamroukhi, Glotin, et Samé 2013). L'entrainement d'un modèle est répété 50 fois car chaque entrainement débouche par construction sur un modèle optimal qui dépend de l'initialisation (Figure 51). L'entrainement est réalisé à l'aide de l'algorithme EM qui réalise une succession d'opération « d'estimation » (les individus sont classés au regard des paramètres du modèle de régression à l'étape donné), et de « maximisation » (les coefficients des modèles sont recalculés pour maximiser la vraisemblance complétée). Parmi tous les entrainements le modèle retenu est celui qui a le critère BIC minimal. Les calculs sont réalisés à l'aide de codes fournis par Allou Samé sur Matlab. Le code a été modifié pour intégrer la capacité à intégrer des variables de régression et améliorer la convergence : l'initialisation des classes est faite 50% du temps aléatoirement, et 50% en classifiant les données de $\boldsymbol{X_1}$ à l'aide de l'algorithme *kmeans*.

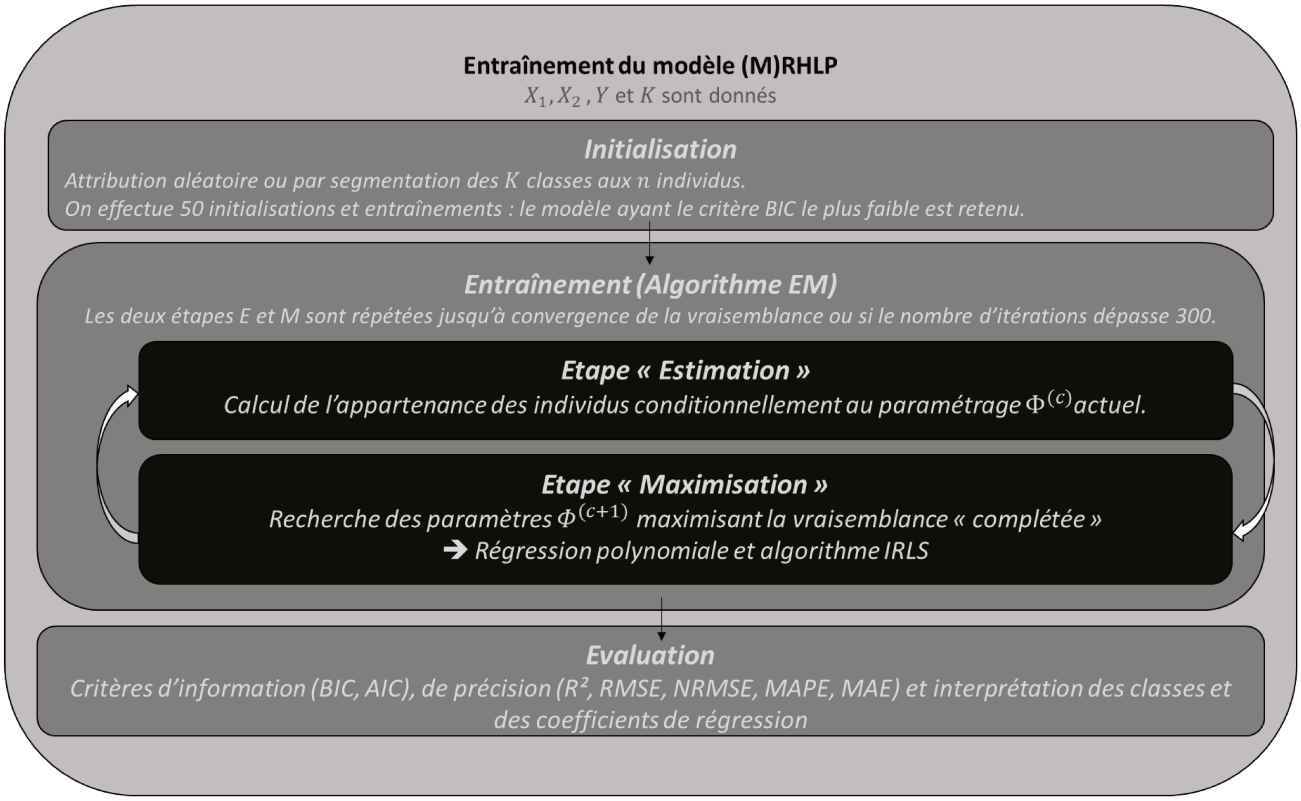


Figure 51 : Procédure d'entrainement d'un modèle RHLP ou MRHLP. Source : Auteur d'après (Chamroukhi, 2013).

3.5 Construction d'un espace synthétique permettant de représenter les situations d'habitation

Les coordonnées $\boldsymbol{X_1}$ des situations d'habitation sont au cœur de la modélisation proposée, puisqu'elles permettent de calculer les probabilités $\pi_k(\boldsymbol{X_1})$ d'appartenance au groupe k. On comprend donc que les hypothèses permettant le calcul de ces coordonnées sont fondamentales dans la construction du modèle. Le critère fondamental pour construire cette matrice de représentation des ménages est qu'elle doit être de telle manière à ce que des ménages « similaires » soit proches au sens de la distance euclidienne dans l'espace $\boldsymbol{X_1}$. Il nous faut en amont définir la « similarité » des situations d'habitation dans l'espace initial. Pour cela, nous avons exploré plusieurs méthodes permettant de définir cela en utilisant des méthodes géométriques (on définit une distance à partir des variables initiales) et des méthodes probabilistes (on définit la densité des voisinages de chacun des points). Une difficulté supplémentaire est la manipulation de variables mixtes. Pour faire cela, plusieurs méthodes ont été explorées :

- L'analyse factorielle de données mixtes permet de calculer des facteurs principaux quantitatifs maximisant l'inertie des données dans l'espace créé. Cette méthode présente deux avantages majeurs qui sont sa simplicité d'utilisation et la stabilité de ses résultats. Aussi, il était attendu que cet algorithme puisse fournir une bonne « séparation » des données (le nuage dans l'espace

synthétique devrait permettre une bonne distinction des contextes résidentiels), mais l'algorithme MRHLP a montré des difficultés à converger ce qui nous a amené à écarter la méthode.

- L'autoencodage des données consiste à calculer un espace synthétique à l'aide d'un réseau de neurones. L'idée est de produire un modèle qui doit reproduire les entrées avec la meilleure précision, et avec une contrainte majeure qui est de passer dans une couche intermédiaire de petite dimension. Par exemple, on demandera au réseau de neurones de produire un estimateur précis de 15 variables (dont 4 variables quantitatives) en passant par un espace à 2 ou 3 dimensions par exemple. On mesure la qualité de l'encodage en mesurant la perte d'information à la sortie du réseau de neurones. Cette méthode a été explorée et a présenté des résultats intéressants mais n'a pas été retenue en raison de l'instabilité des résultats.
- La méthode t-SNE (pour *t-distributed stochastic neighbor embedding* en anglais) est un algorithme permettant de calculer un espace synthétique à 2 ou 3 dimensions à partir d'une représentation probabiliste des distances. L'algorithme cherche à construire un espace où deux ménages proches (resp. éloignés) dans l'espace de départ sont proches (resp. loins) dans l'espace synthétique. Le critère de construction compare des densités de probabilité modélisant la probabilité des voisinages dans les espaces d'arrivée et de départ. Cette méthode permet de fournir des espaces synthétiques très intéressants pour la visualisation des données mais si l'algorithme permet de restituer les voisinages, les distances plus importantes ne sont pas représentatives dans l'espace synthétique, rendant difficile son utilisation. Nous avons donc également écarté cette méthode.
- Une dernière méthode que nous avons testée consiste à calculer les distances de Gower entre les situations d'habitation. Ensuite, la méthode MDS non métrique (pour l'anglais *Multidimensional Scaling*) place des individus dans un espace synthétique de dimension p et optimise leur place pour minimiser une fonction de *stress.* Cette fonction vise à ce que l'ordre des voisinages dans l'espace initial soit respecté dans l'espace synthétique calculé. Cette méthode a présenté des résultats interprétables, avec une bonne précision de régression et une bonne stabilité. Nous présentons donc les résultats obtenus par cette méthode.

La valeur de p est choisie égale à 6 car les gains en stress pour des valeurs supérieures de p étaient moins importants et il était difficile d'interpréter les axes suivants (Figure 52).

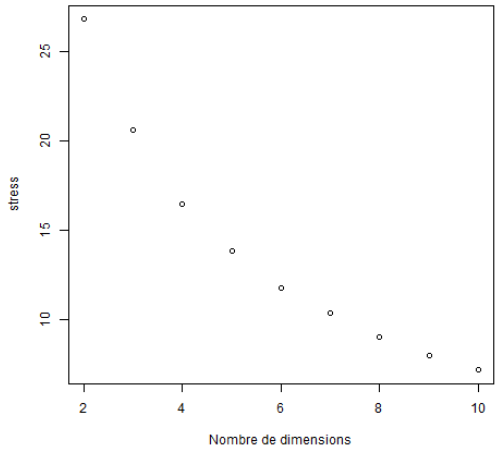


Figure 52 : Valeurs de la fonction de stress pour différentes de la dimension p de l'espace synthétique. Source : Auteur.

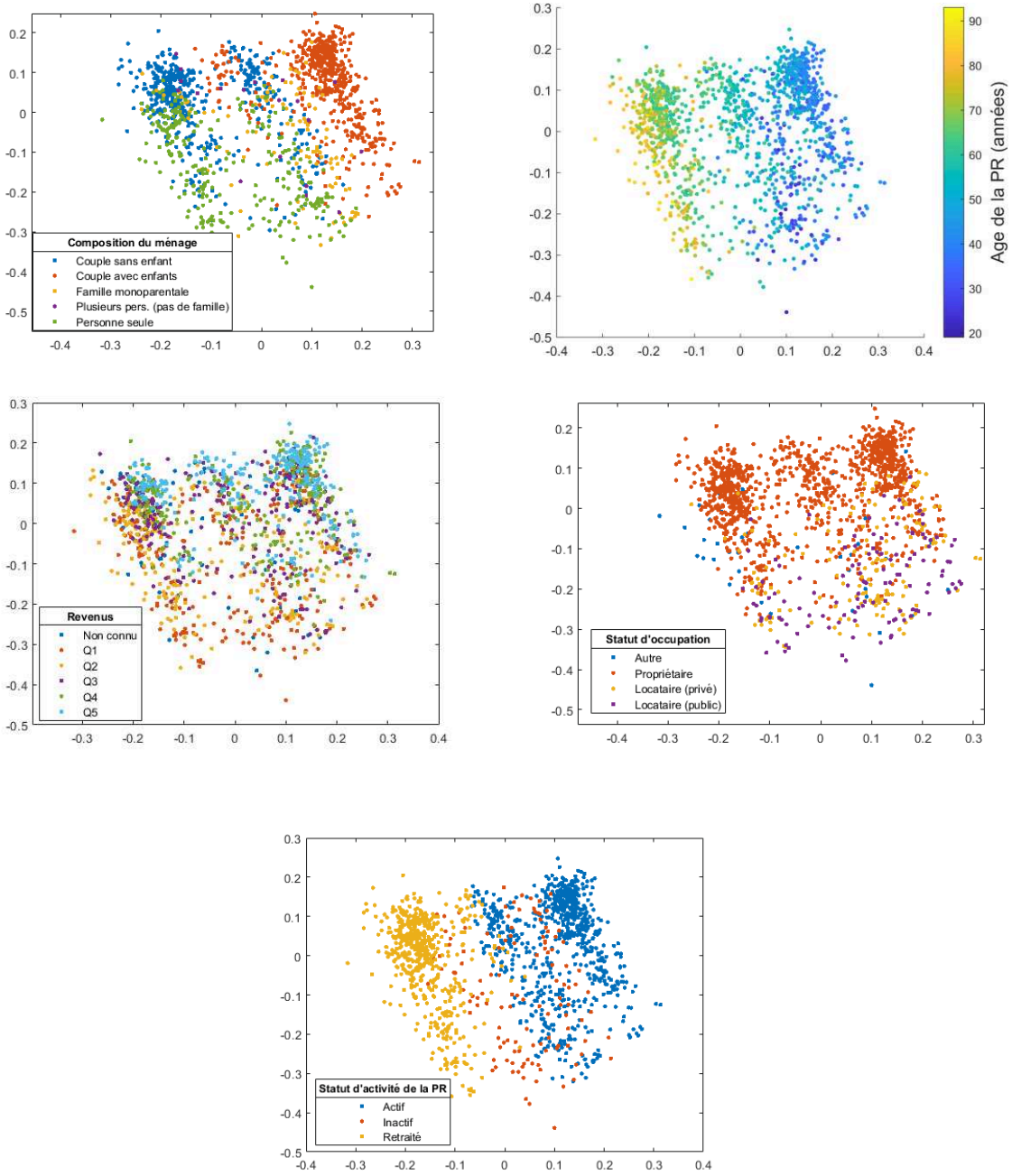
Il devient alors possible de représenter une cartographie des ménages et des logements. Ceux-ci sont représentés selon les dimensions 1 et 2 sur la Figure 53. Les autres dimensions ne sont pas représentées ici mais on donne ici un résumé de l'interprétation qualitative des graphes.

- La dimension 1 est liée à l'âge de la PR et la composition du ménage
- La dimension 2 est liée à la surface, au type de logement, au type de chauffage, au statut d'occupation et au revenu.
- La dimension 3 est liée au type de logement et au type de chauffage.
- La dimension 4 est liée principalement au niveau de revenu
- La dimension 5 est liée au type de ménage et au type de logement
- La dimension 6 est liée au type de chauffage.

On remarque au passage que la variable FEC est très liée à la dimension 2, ce qui est cohérent avec le fait que le modèle FEC est liée à la surface et au type de logement.

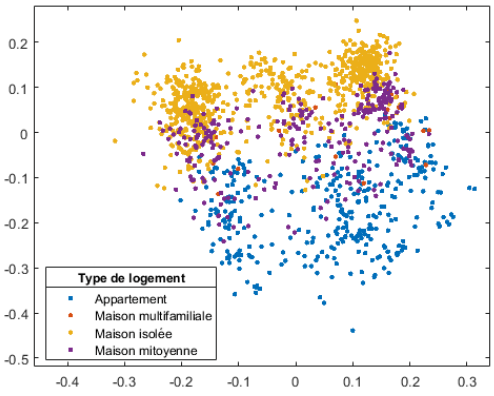
Caractéristiques des ménages

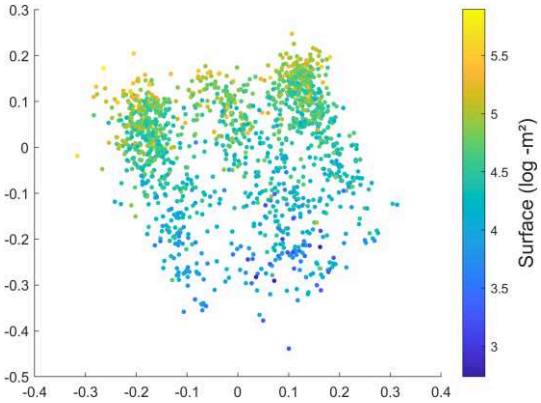
Projection dans les dimensions 1 et 2 de l'espace calculé par MDS

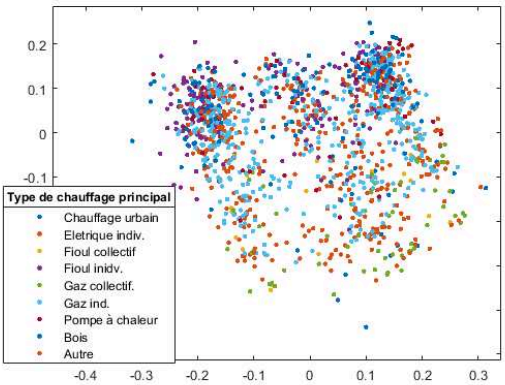


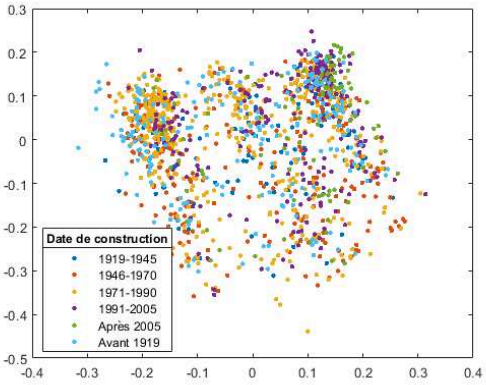
Caractéristiques des logements

Projection dans les dimensions 1 et 2 de l'espace calculé par MDS









Consommation énergétique

Projection dans les dimensions 1 et 2 de l'espace calculé par MDS

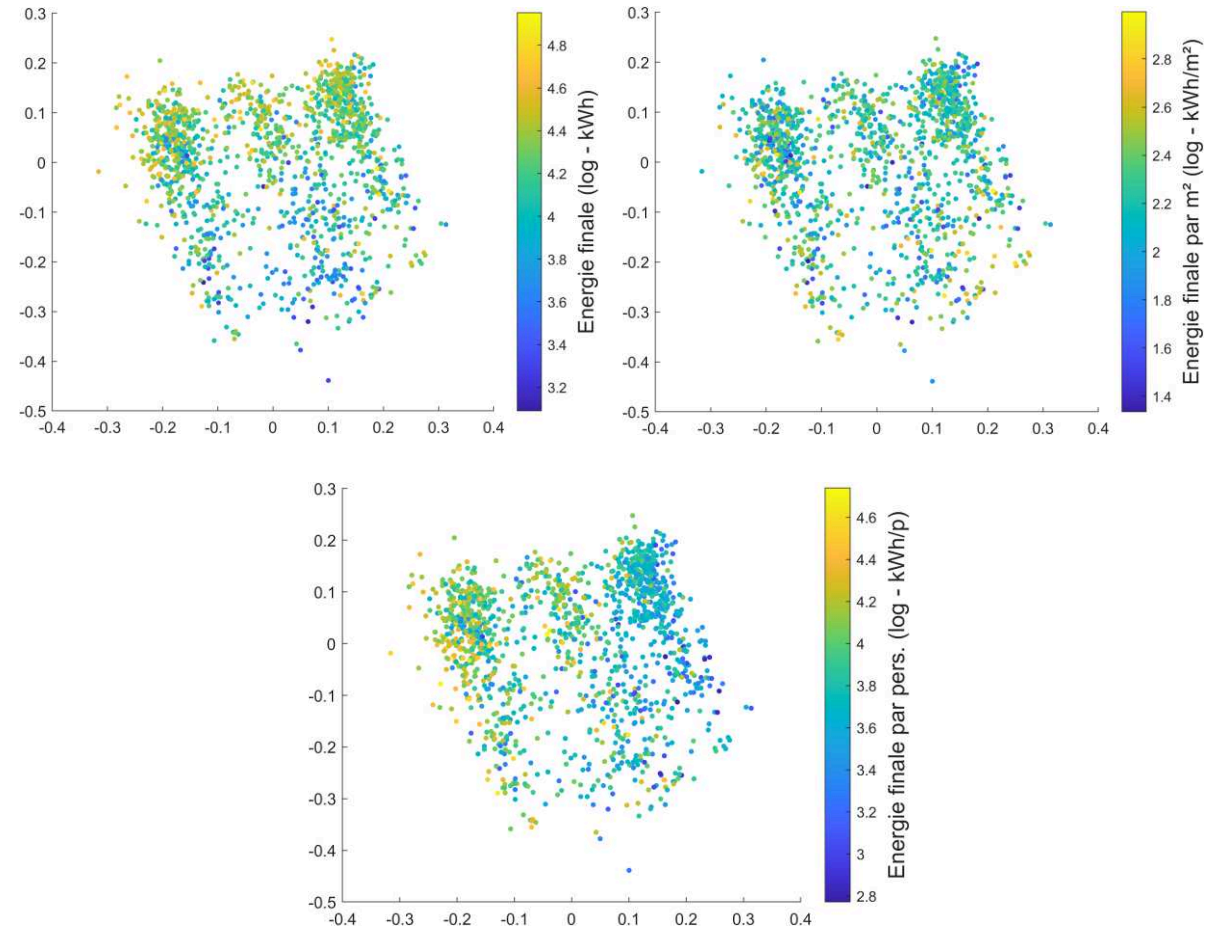


Figure 53 : Représentation des ménages et des logements dans l'espace synthétique des variables X1 calculé par la méthode MDS non métrique. Les figures sont identiques au code couleur près : les cartes permettent de visualiser la distribution de plusieurs variables sur les dimensions 1 et 2 de l'espace synthétique. Source : Auteur après calculs sur la base PHEBUS.

3.6 Construction des variables synthétiques de comportement

Après le choix de l'espace de représentation et de partitionnement $\boldsymbol{X_1}$, Une seconde hypothèse de modélisation est celle du choix des variables composant la matrice de régression $\boldsymbol{X_2}$. Dans la perspective de cette thèse, nous avons initialement proposé d'intégrer les variables explicatives de la consommation d'énergie à savoir les facteurs extensifs du bâti (surface, qualité de l'isolation) et les facteurs comportementaux.

Note méthodologique importante sur l'intégration des variables de comportement

L'intégration des variables de comportement dans le modèle RHLP en tant que variables de régression n'a pas permis de fournir des modèles avec des classes suffisamment homogènes. Statistiquement, cela peut être lié au fait que les coefficients de régression associés à chacun des clusters sont relativement proches. Une autre hypothèse, alimentée par nos expériences, est que les coefficients associés aux variables modélisant les pratiques sont faibles devant les coefficients associés aux variables associées à la surface et à l'isolation. L'asymétrie des poids statistiques entre les variables empêche le modèle de converger. Pour poursuivre les investigations, nous avons choisi de ne conserver que la surface et le niveau d'isolation comme variables de régression et d'utiliser les variables décrivant les pratiques comme des variables supplémentaires pour l'analyse du modèle. Cette approche, cohérente avec le cadre de modélisation, ne permet pas de quantifier les effets des pratiques sur le modèle (c'est la principale limite), mais il permet de vérifier la cohérence de notre approche. Nous présentons les résultats associés à cette hypothèse.

Pour intégrer les comportements, il faut proposer des variables quantitatives mais qui soient également interprétables. En nous basant sur les calculs de la partie précédente, nous proposons de réaliser un travail de classification à l'aide de la méthode *ClustOfVar*. En suivant la méthodologie de classification des variables présentée dans la partie Méthodologie p.86), nous construisons 8 variables synthétiques à partir des variables de comportement extraites de la base PHEBUS. Un résumé des résultats est fourni en annexe (voir p.237). On constate que les variables de comportements identifiées sont similaires à celles calculées sur la base ENERGIHAB (équipement occupation moyenne du logement demande en chauffage, comportements de restriction). Pour mieux connaitre le lien entre les VS et les indicateurs de consommation (la consommation en énergie finale totale FEC, par personne FECP et par mètre carré FECM2) nous avons effectué une régression simple (Tableau 25). La régression de ces variables permet d'observer une liaison statistique entre FEC et toutes les variables à l'exception de VS2 (modélisant le nombre d'équipements efficaces achetés) et VS8 (modélisant l'intensité des gestes de régulation thermique). Celles-ci s'avèrent en revanche liées aux indicateurs FECP et FECM2.

Tableau 25 : Inventaire des effets simples calculés entre les 3 indicateurs de consommations et les variables synthétiques de comportements, calculés selon la méthode ClustOfVar. n.s : non significatif. Source : Auteur, d'après les données PHEBUS.

	VS1 (Equipement)	VS2 (Efficacité des équipements)	VS3 *(Usage de l'ECS)*	VS4 *(Demande en chauffage)*	VS5 *(Inoccupations longues du logement)*	VS6 *(Présence au logement)*	VS7 *(Comportements de restriction)*	VS8 *(Niveau de régulation)*
FEC	-0,07 (***)	n.s.	0,02 (**)	0,05 (***)	-0,01 (**)	0,04 (***)	0,02 (***)	n.s
FECM2	n.s	n.s	n.s	0,05 (***)	-0,03 (***)	0,01 (*)	n.s	-0,08 (**)
FECP	n.s	-0,03 (***)	-0,03 (***)	0,04 (***)	n.s	0,02 (**)	0,01 (*)	n.s

Distribution des valeurs prises par les variables synthétiques de comportement dans l'espace MDS (axes 1 et 2)

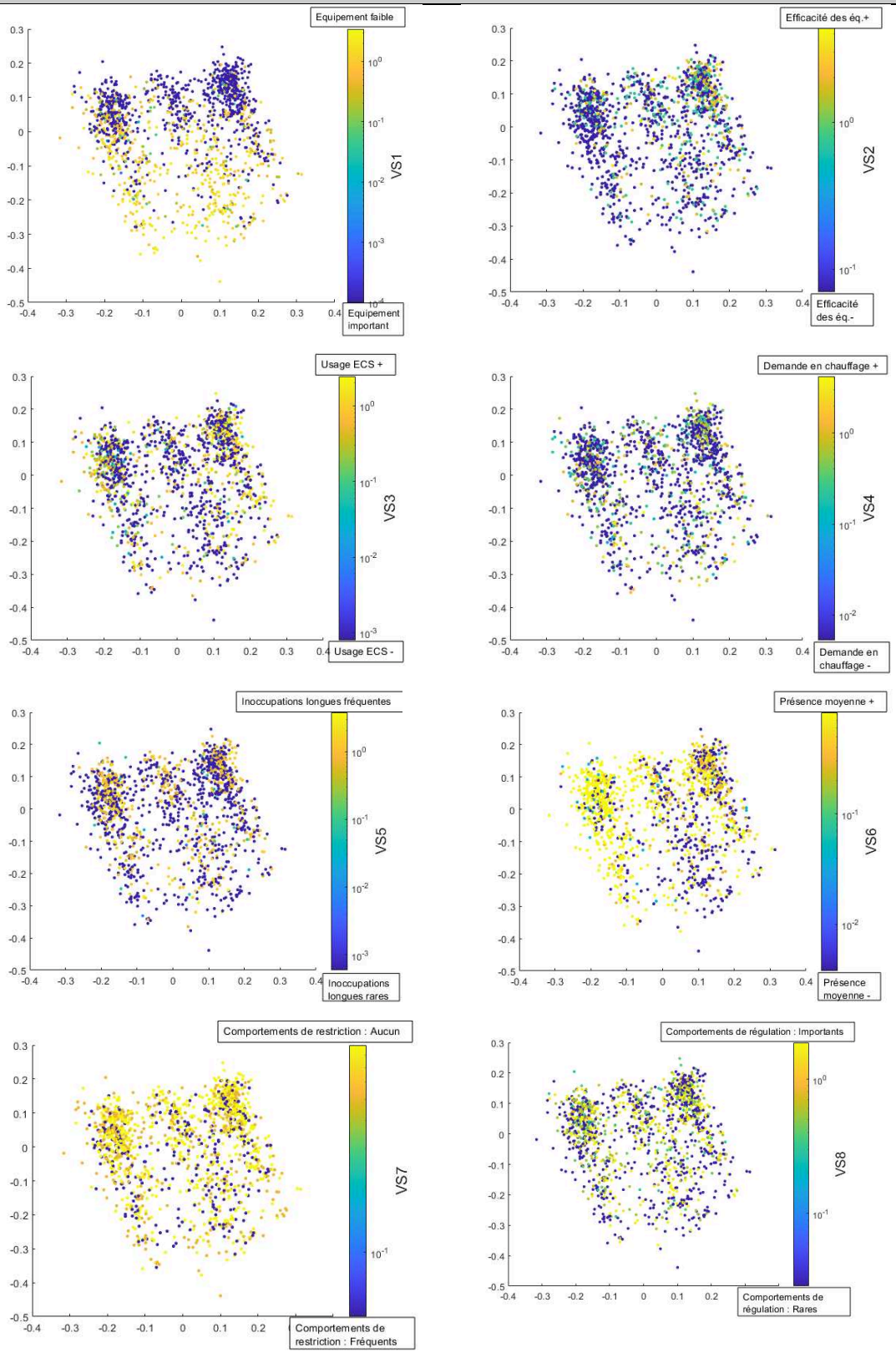


Figure 54 : Distribution des VS dans l'espace synthétique (dimensions 1 et 2). Les couleurs sont données en échelle logarithmique pour faciliter la lecture. Source : Auteur après calculs sur la base PHEBUS.

Il est également intéressant d'observer la distribution des VS dans l'espace $\boldsymbol{X_1}$. On peut remarquer que les distribution de VS1 est très liée à cet espace : les ménages les plus équipés (valeurs les plus faibles de VS1) sont plutôt sur la partie supérieure (Figure 54). Aussi, l'usage de l'ECS (VS3) est plus important sur la partie située à droite du nuage de points et les inoccupations longues du logement (VS5) sont plus importantes dans la partie supérieure. La présence au logement (VS6) est plus importante sur la partie gauche du nuage de points. Cependant si la plupart des VS semblent liée à certaines localisations dans cet espace (ce qui rejoint l'idée défendue dans le chapitre 2 d'un lien entre contextes résidentiels et pratiques domestiques), on peut voir que les variables VS4 et VS2 semblent moins liées à cet espace.

Si le souhait initial était de pouvoir intégrer les variables VS dans le modèle MRHLP, nous nous sommes toutefois rendus compte lors du processus de sélection des variables qu'en dépit du lien statistique avéré entre les variables VS et les indicateurs de consommation FEC, FECM2 et FECP, les valeurs des effets de chacune des variables était trop faible par rapport aux effets de la surface et du DPE. Cette difficulté nous a amené à considérer l'opportunité de n'utiliser alors les variables synthétiques décrivant les pratiques comme des variables supplémentaires permettant de décrire les types construits par la méthode MRHLP. Nous décrivons dans les lignes qui suivent la construction des différents modèles MRHLP.

3.7 Construction d'un modèle pour chaque indicateur de consommation

3.7.1 Modèle 1 : modélisation de l'indicateur FEC

Sélection du modèle

Le modèle 1 est construit selon la méthodologie présentée dans la section précédente. On trace sur la figure suivante l'évolution des critères d'information (probabilistes) et des performances de prédiction en fonction du nombre de classes K. L'entraînement du modèle, c'est-à-dire le calcul du partitionnement et des paramètres associés n'est pas déterministe : il dépend essentiellement de l'initialisation. C'est pourquoi nous réalisons le calcul 50 fois pour chaque valeur de K. Nous traçons ainsi les évolutions tendancielles ainsi que l'incertitude calculée (écart-type) sur ces 50 calculs (Figure 55).

On observe sur cette figure un accroissement des performances d'estimation (augmentation du R^2) sur les données d'entrainement et de test pour K variant de 1 à 7 puis un décrochage entre les deux ce qui marque un effet de surapprentissage. En parallèle les indicateurs d'information (BIC et AIC) décroissent ensemble jusqu'à $K = 3$. Ensuite, le critère BIC croît alors que le critère AIC se stabilise autour d'une asymptote. Ces critères informationnels invitent ainsi à sélectionner un modèle pour $K = 3\ ou\ 4$, en considérant les marges d'erreur. En termes d'erreur, on observe que l'erreur quadratique diminue en moyenne jusqu'à $K = 3$. Pour $K = 4$ et au-delà on observe un risque de sur-apprentissage : le modèle offre de bonnes performances sur les données utilisées pour l'apprentissage mais des performances

dégradées sur les données de test. En analysant les types construits pour ces deux valeurs de K, on choisit finalement $K = 3$.

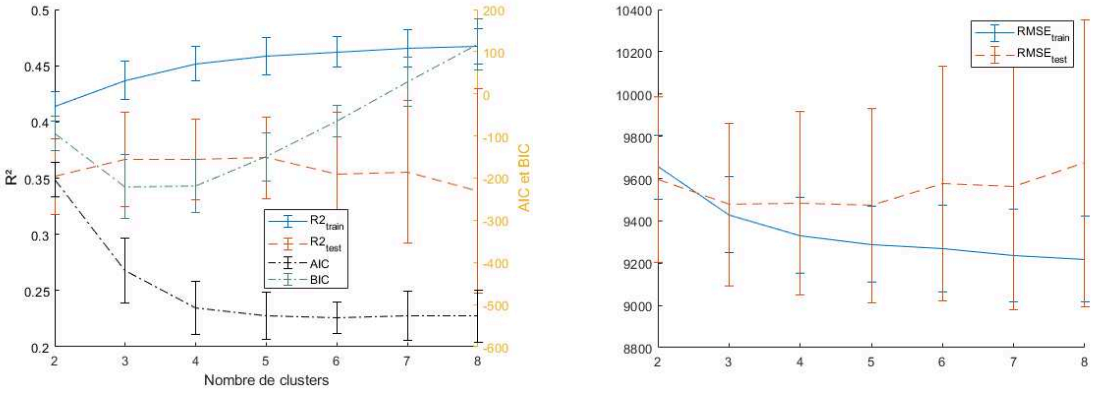


Figure 55 : Evolution des critères d'information (AIC et BIC) et de précision (R²) (à gauche) et de l'erreur quadratique (à droite) en fonction du nombre de clusters choisis pour l'algorithme MRHLP. On donne les critères de performance sur les ensembles d'entraînement (train) et de test (test) pour repérer un éventuel surapprentissage du modèle. Les barres d'erreurs sont tracées pour représenter l'écart-type des valeurs calculées après entrainement sur 50 ensembles de données entrainement/test. Source : Auteur après calculs sur la base PHEBUS.

La validation du modèle retenu est faite en plusieurs étapes. Premièrement nous vérifions qu'il existe une bonne séparation des données en observant les 3 distributions des coefficients π_{ik} qui manifestent l'appartenance de la ligne i au cluster k. On observe que les valeurs sont bien séparées (distribution en forme de U) ce qui témoigne d'une bonne séparation des données (Figure 52).

Répartition des lignes de la base d'entraînement

Cluster 1 : 748 lignes
Cluster 2 : 536 lignes
Cluster 3 : 105 lignes

Performances du modèle retenu

$AIC = -481$
$R^2_{train} = 0{,}44$
$R^2_{test} = 0{,}39$

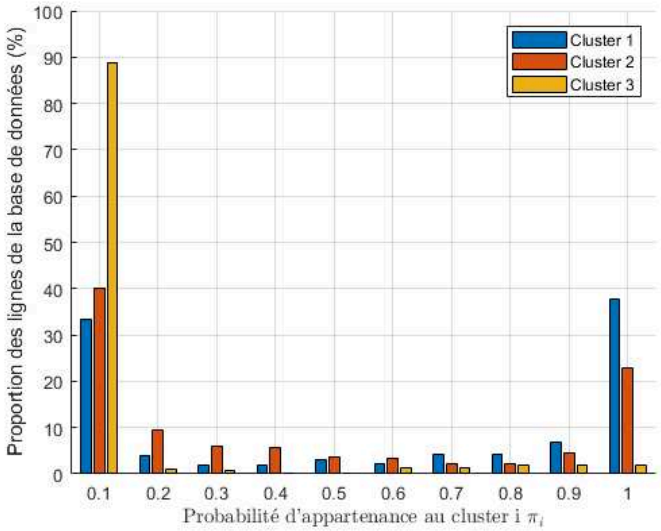


Figure 56 : Performance finales du modèle de l'indicateur FEC avec K=3. Sur le graphe à droite on observe les distributions des coefficients π_{ik}. Aide de lecture : 38% des lignes de la base de données (décrivant des ménages et des logements) ont une probabilité d'appartenance au cluster 1 située entre 0,9 et 1 (soit entre 90 et 100%). Source : Auteur après calculs sur la base PHEBUS.

Ensuite, on observe la performance de l'estimateur ainsi construit. Il propose un R² sur l'ensemble de test de 0,39 en moyenne ce qui est légèrement inférieur à ce qui a été obtenu à l'aide du modèle multilinéaire (0,45). Le fait que les estimations fournies par ce modèle ne soient pas (trop) fausses,

relativement au modèle de référence permet ainsi de ne pas écarter le modèle et de poursuivre son analyse. Aussi, il est intéressant de constater que le partitionnement calculé offre une performance prédictive à peine inférieure au modèle de régression multilinéaire calculé à la partie précédente, alors même que seuls deux coefficients de régression utilisés, en complément de l'utilisation de variables de partitionnement.

Analyse du modèle optimal retenu

Présentation des clusters

Le modèle retenu est un modèle à 3 classes. Nous présentons les types, les performances d'estimation et les coefficients des modèles de régression calculés. On peut observer une bonne séparation des clusters en les représentant dans l'espace synthétique MDS (Figure 57).

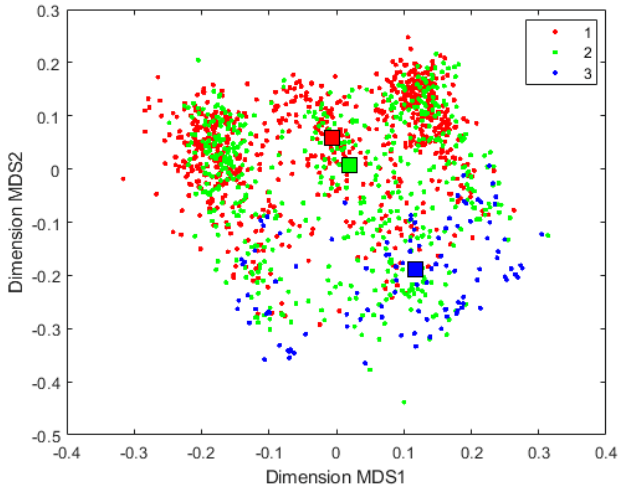


Figure 57 : Répartition des ménages et des logements dans les 3 clusters identifiés par le modèle MRHLP. Source : Auteur après calculés sur la base PHEBUS.

En analysant les distributions des variables caractéristiques des logements et des ménages par cluster (Figure 58), on peut caractériser les profils typiques des situations d'habitation de chaque groupe. En particulier, on observe que le partitionnement opéré par ce modèle distingue fondamentalement les types de logement. Le groupe n°1 regroupe plutôt des maisons individuelles chauffées au gaz ou au fioul. Ces logements sont aussi souvent les plus anciens de l'échantillon. Le second groupe contient quant à lui des maisons individuelles chauffés à l'électricité le plus souvent et qui sont tendanciellement plus récentes. Le groupe n°3 enfin est composé uniquement d'appartements, bénéficiant souvent d'un chauffage collectif. Le partitionnement proposé ici ne différencie pas les types de ménage.

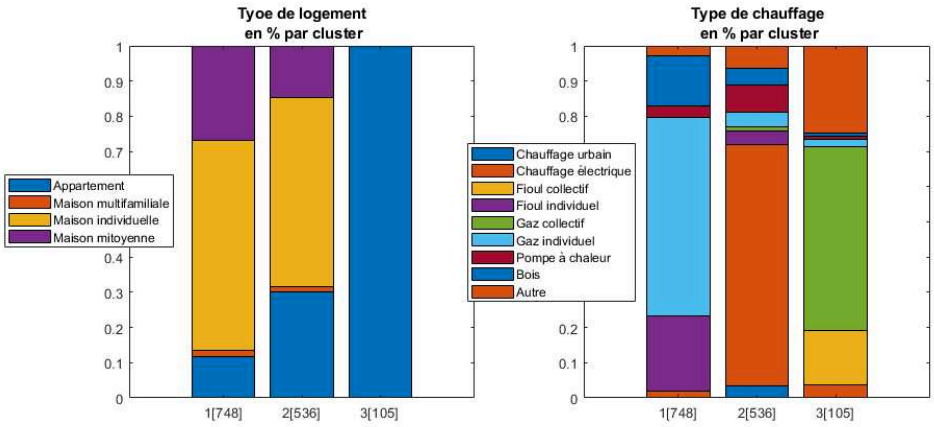


Figure 58 : Illustration graphique de la distribution des variables caractérisant le ménage, ventilés par cluster. Source : Auteur après calculs sur la base PHEBUS.

Analyse des coefficients

On rappelle que le modèle repose sur la sélection suivante des variables :

- $\boldsymbol{X_1 = [1, MDS, MDS^2]}$
- $\boldsymbol{X_2 = [1, log(SURF), log(DPE)]}$
- $\boldsymbol{Y = log(FEC)}$

Où $\boldsymbol{MDS}$ désigne la matrice composée des 6 vecteurs contenant les positions des ménages dans l'espace synthétique MDS. On peut alors étudier les coefficients du modèle construit pour étudier la sensibilité relative des clusters aux différentes variables de régression (Figure 59).

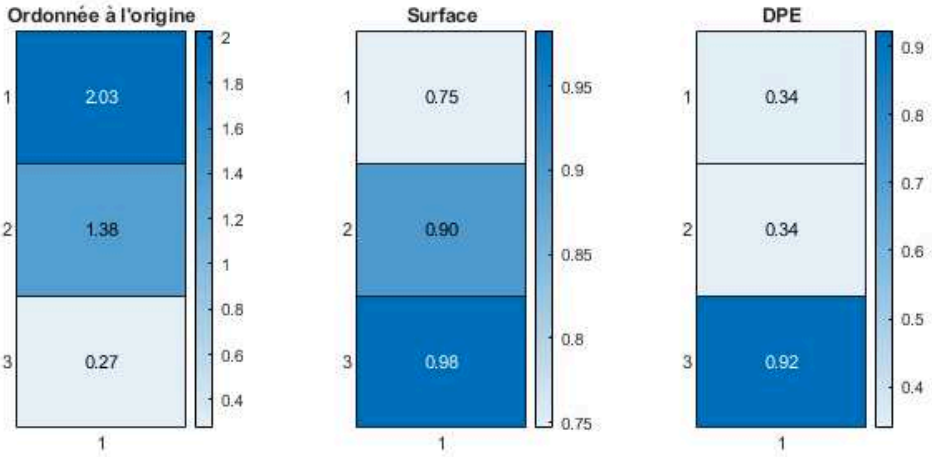


Figure 59 : Valeur des coefficients de régression pour chacun des types. Les types sont en ligne et les coefficients en colonne. Source : Auteur après calculs sur la base PHEBUS.

On observe sur cette figure que les coefficients diffèrent selon les clusters. Le groupe 3, constitué de ménages vivant en appartements ont une sensibilité particulièrement élevée à la surface et au DPE. A contrario, le cluster 1 offre une sensibilité moindre à la surface et au DPE. Le cluster 2 quant à lui offre une sensibilité élevée à la surface, et réduite au DPE.

Cette observation peut être mise en regard avec les niveaux de consommation observés pour chacun de ces clusters. Les clusters 1 et 3 ont des niveaux de consommations relativement proches et différents du cluster 2.

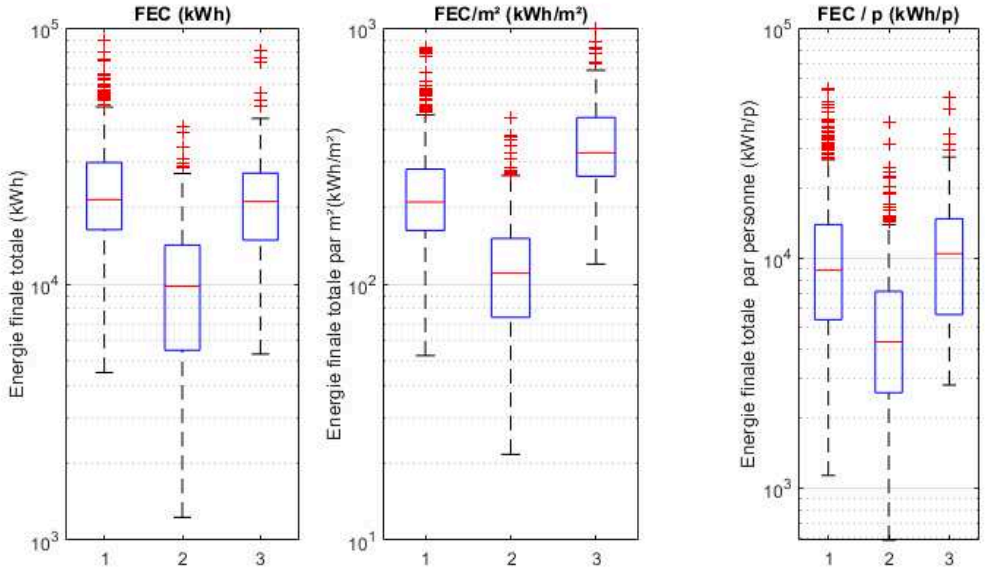


Figure 60 : Distributions des indicateurs de consommation pour chacun des clusters. Les consommations sont tracées sur une échelle logarithmique. Source : Auteur après calculs sur la base PHEBUS.

La proximité des consommations et la différence significative des modèles sur les deux groupes 1 et 3 suggèrent que les explications de niveaux de consommations sont différentes pour chacun des types (Figure 60). On peut alors, en plus des caractéristiques des logements et des ménages étudier les pratiques résidentielles éventuellement liées à chacun des types. En particulier, en observant les niveaux moyens des variables synthétiques pour chacune des classes on remarque que :

- Groupe 1 : les ménages sont réputés plus équipés que la moyenne, peu présents au domicile, et ayant peu de gestes de régulation.
- Groupe 2 : les ménages indiquent en moyenne avoir une demande en chauffage plutôt basse, avec des gestes de régulation, voire de restriction importants.
- Groupe 3 : les ménages rattachés à ce type ont un niveau d'équipement plus bas que la moyenne mais une demande en chauffage élevée accompagnés toutefois de gestes de régulation.

On constate alors en croisant les types construits avec les indicateurs de consommation et les variables de comportement que le partitionnement opéré distingue non seulement des types de logement et de chauffage mais aussi des pratiques de chauffage et des niveaux d'équipement. Ces groupes ne paraissent toutefois pas assez homogènes pour pouvoir décrire suffisamment précisément les logiques de consommations latentes. Nous proposons dans la suite de suivre la même méthodologie pour construire des partitionnements et des modèles de consommations des indicateurs de FECM2 et FECP. En effet,

ces deux autres indicateurs sont deux autre manières de « compter », de quantifier les services énergétiques domestiques des ménages, en les ramenant à la taille du logement ou du ménage.

3.7.2 Modèle 2 : modélisation de l'indicateur FECM2

On suit dans cette partie une démarche similaire à la partie précédente pour construire un modèle MRHLP permettant de modéliser le logarithme de l'indicateur FECM2.

Le modèle construit repose sur la sélection suivante des variables :

- $\boldsymbol{X_1} = [\mathbf{1}, \boldsymbol{MDS}, \boldsymbol{MDS^2}]$
- $\boldsymbol{X_2} = [\mathbf{1}, \boldsymbol{log}(\boldsymbol{SURF}), \boldsymbol{log}(\boldsymbol{DPE})]$
- $\boldsymbol{Y} = log(\boldsymbol{FECM2})$

Où $\boldsymbol{MDS}$ désigne la matrice composée des 6 vecteurs contenant les positions des ménages dans l'espaces synthétique MDS.

Sélection du modèle

De manière similaire au paragraphe précédent, on sélectionne les variables explicatives puis on explore l'évolution des critères d'information et de précision pour différentes valeurs de K. Par un raisonnement similaire, on voit qu'il y a surapprentissage pour K supérieur à 8. Par ailleurs le critère BIC indique un optimum pour $K = 4$. Le critère AIC évolue selon une asymptote horizontale pour K supérieur à 4. En étudiant le modèle fourni pour $K = 4$ on observe que les classes sont différentes en composition et dans les coefficients de régression. On retient donc $K = 4$.

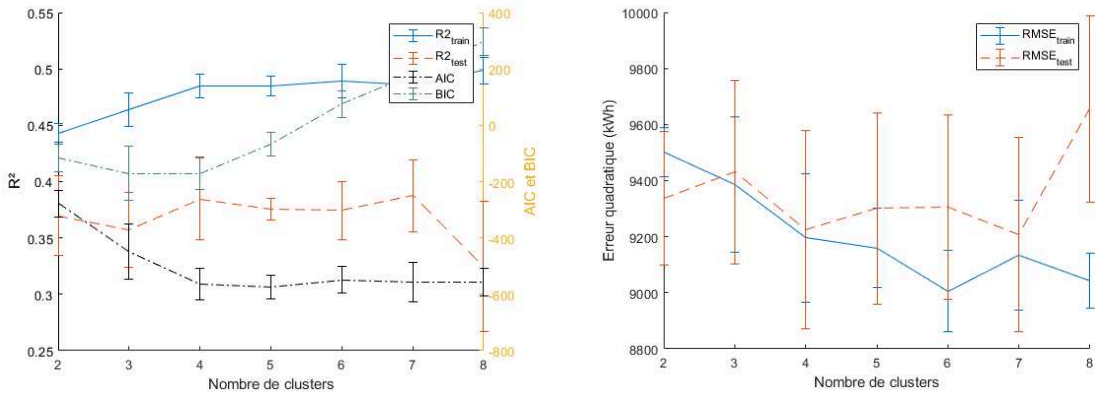


Figure 61 : Evolutions des critères d'information (AIC et BIC) et de précision (R² et RMSE) pour le modèle MRHLP de l'indicateur FECM2. Source : Auteur après calculs sur la base PHEBUS.

Analyse du modèle retenu

Le modèle calculé possède 4 classes. On observe un séparation correcte des différentes classes (Figure 62) ce qui permet de valider le partitionnement. Par ailleurs les performances de régressions sur l'ensemble de test ($R^2{}_{test} = 43\%$) sont similaires à ce qui a été observé dans la partie précédente à

l'aide d'un modèle multilinéaire simple ($R^2{}_{test} = 42\%$) . Comme dans le cas précédent, cette performance d'estimation est obtenue avec quatre modèles (un pour chaque groupe) contenant chacun deux coefficients de régression (plus un coefficient pour l'ordonnée à l'origine).

Répartition des lignes de la base d'entraînement

Cluster 1 : 748 lignes
Cluster 2 : 331 lignes
Cluster 3 : 196 lignes
Cluster 4 : 114 lignes

Performances du modèle retenu

$AIC = -552$
$R^2{}_{train} = 0{,}48$
$R^2{}_{test} = 0{,}43$

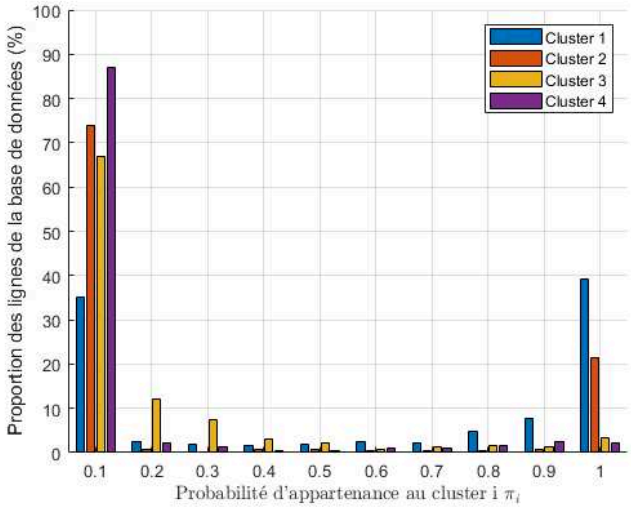


Figure 62 : Eléments descriptifs du modèle retenu pour l'indicateur FECM2. La figure de droite décrit la distribution des paramètres. Aide de lecture : 35% des lignes de la base de données (décrivant des ménages et des logements) ont une probabilité d'appartenance au cluster 1 située entre 0 et 0,1 (soit entre 0 et 10%). Source : Auteur après calculs sur la base PHEBUS.

L'analyse des caractéristiques des logements, des ménages, des consommations et des comportements permet de décrire les quatre classes (Figure 63) :

- Classe 1 : dans cette classe on retrouve essentiellement des ménages aux revenus élevés, occupant des maisons individuelles, chauffées au gaz. Les ménages ont un équipement très important et présentent des consommations en énergie finale importantes.
- Classe 2 : dans cette classe, on retrouve des situations d'habitation similaires au premier segment. La différence principale est dans le mode de chauffage qui est plutôt électrique. On note aussi que les consommations énergétiques sont plutôt dans la moyenne de l'échantillon.
- Classe 3 : cette classe comprend des ménages composés de personnes plutôt âgées et vivant seules ou en couple et sans enfant. Plutôt présents au domicile, et ayant des comportements de restriction ces ménages vivent plutôt dans des logements collectifs.
- Classe 4 : cette classe se caractérise par des ménages occupant des logements collectifs, et ayant une demande en chauffage plutôt élevée.

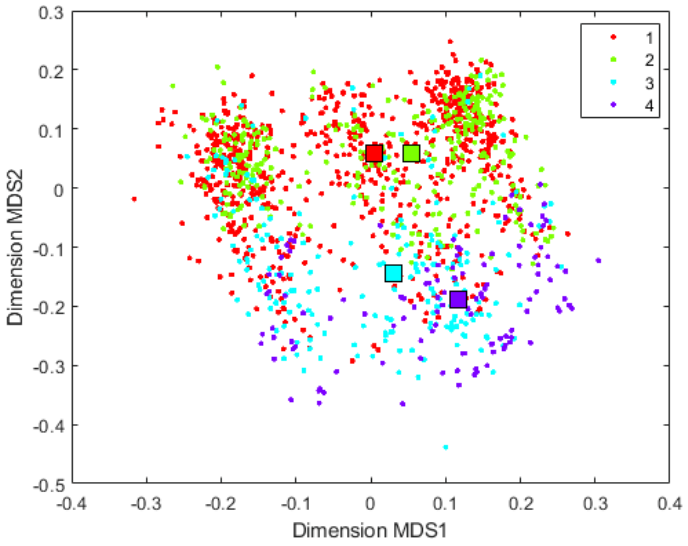


Figure 63 : Représentation des situations d'habitation dans l'espace (MDS1, MDS2). Un code couleur permet de distinguer les quatre classes construits lors de la modélisation de l'indicateur FECM2. Source : Auteur après calculs sur la base PHEBUS.

Enfin, l'analyse des coefficients de régression par cluster est aussi intéressante (Figure 64). On observe globalement une liaison non nulle mais faible et négative entre l'indicateur FECM2 et la surface. D'après ce modèle, cela signifierait que à situation d'habitation donnée, un ajout de surface engendrerait un surplus de consommation d'énergie final presque proportionnel à la surface ajoutée. Cela est particulièrement vrai pour le cluster 4 et un peu moins pour les maisons individuelles (cluster 1 et 2). En revanche, ce lien n'est plus avéré pour le cluster 3 qui regroupe plutôt des ménages âgés et sans enfants. Dans ce cluster, la relation entre FECM2 et la surface est négative ce qui témoigne du fait que les services énergétiques sont restreints dans cette situation d'habitation et l'ajout de surface n'engendre que peu de nouveaux services énergétiques. Cette remarque est en accord avec le diagnostic fait du fait que ces ménages ont des comportements de régulation et de restriction plus importants que la moyenne. En termes de DPE, la dynamique est légèrement différente puisqu'on observe, en accord avec les calculs effectués dans la partie 2 de ce chapitre que le DPE a une influence plus forte sur la consommation FECM2 dans le cas de logements collectifs.

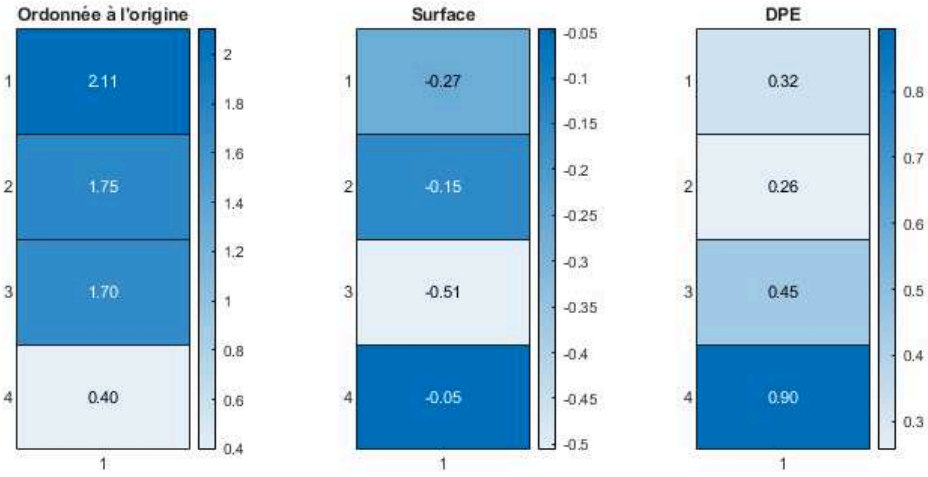


Figure 64 : Inventaire des coefficients de régression calculés pour chaque cluster. Chaque table renseigne les coefficients de régression pour une variable. Chaque ligne contient le coefficient de régression d'un cluster. Par exemple, le cluster 1 a un modèle de régression de l'indicateur avec un coefficient de régression linéaire de -0,27 avec le logarithme de la surface. Source: Auteur après calculs sur la base PHEBUS.

La modélisation de l'indicateur FECM2 a abouti sur un partitionnement relativement proche du partitionnement précédent, en différenciant toutefois une classe qui regroupe des personnes âgées vivant dans des logements collectifs ou de petits logements individuels et qui adoptent des gestes de régulation importants. Dans la partie qui suit, nous entreprenons la modélisation de l'indicateur FECP. Ce dernier indicateur permet de compter les services énergétiques des ménages en mettant l'accent sur leur mutualisation.

3.7.3 Modèle 3 : modélisation de l'indicateur FECP

Sélection du modèle

De manière similaire on recherche une valeur de K optimale de manière à minimiser le critère AIC, ainsi que l'erreur quadratique du modèle MRHLP sur l'ensemble de test.

Le modèle construit repose sur la sélection suivante des variables :

- $\boldsymbol{X_1} = [\mathbf{1}, \boldsymbol{MDS}, \boldsymbol{MDS^2}]$
- $\boldsymbol{X_2} = [\mathbf{1}, \boldsymbol{log}(\boldsymbol{SURF}), \boldsymbol{log}(\boldsymbol{DPE})]$
- $\boldsymbol{Y} = log(\boldsymbol{FECP})$

L'entrainement du modèle de l'indicateur FECP pour K variant de 2 à 9 montre permet d'observer l'évolution des indicateurs d'information et d'erreur avec le nombre de clusters (Figure 65). On observe que le R^2 croit significativement et l'erreur quadratique décroit significativement jusqu'à $K = 5$. Parallèlement l'indicateur BIC est uniquement croissant mais l'indicateur AIC observe un minium pour $K = 5$ également. L'analyse des clusters s'étant révélée satisfaisante pour $K = 5$ nous retenons ce paramétrage.

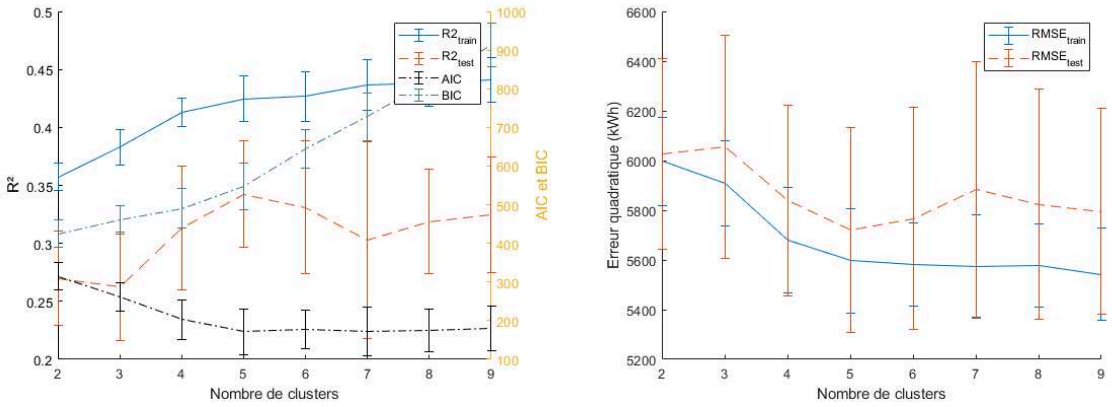


Figure 65 : Evolution des critères d'information (AIC, BIC) et d'erreur (R² et RMSE) avec le nombre de clusters K retenu pour la modélisation de l'indicateur FECP. Source : Auteur après calculs sur la base PHEBUS.

Analyse du modèle optimal

Dans ce cas nous retenons un modèle à 5 classes. Celui-ci présente des performances d'estimation en dessous des performances du modèle multilinéaire (RMSE = 5.3 MWh/p contre 4.5 MWh/p dans ce cas). Par ailleurs, on observe que les paramètres π_{ik} ne sont pas souvent égaux à 0 ou 1 (Figure 66), ce qui marque un partitionnement de moindre qualité. On observe que les clusters 2 et 1 et les clusters 3 et 4 sont relativement proches deux à deux. Pour comprendre cela, on peut remarquer que la modélisation du logarithme de la consommation par personne dépend largement de la variable « surface par personne » qui n'a pas été sélectionné pour ce modèle comme variable de régression. Le modèle construit est ainsi un estimateur moyen comparativement au modèle multilinéaire.

Répartition des lignes de la base d'entraînement

Cluster 1 : 237 lignes
Cluster 2 : 192 lignes
Cluster 3 : 421 lignes
Cluster 4 : 487 lignes
Cluster 5 : 52 lignes

Performances du modèle retenu

$$AIC = 74$$
$$R^2{}_{train} = 0{,}42$$
$$R^2{}_{test} = 0{,}34$$

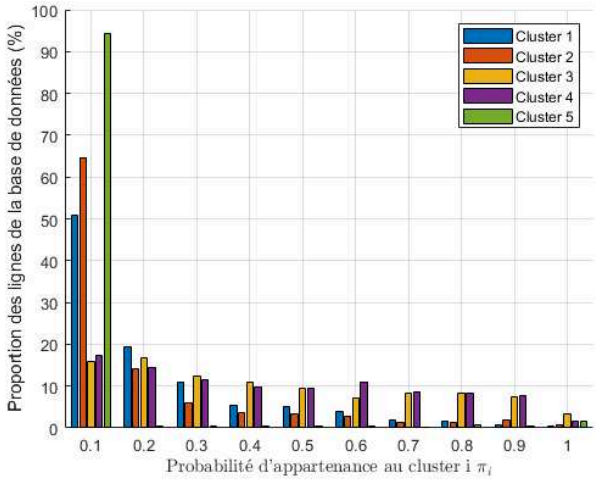


Figure 66 : Position et partitionnement des ménages dans l'espace MDS après calcul du modèle MRHLP pour l'indicateur FECP. Aide de lecture : 65% des lignes de la base de données (décrivant des ménages et des logements) ont une probabilité d'appartenance au cluster 2 située entre 0 et 0,1 (soit entre 0 et 10%). Source : Auteur.

On peut décrire chacun des clusters en croisant les caractéristiques des logements, des ménages, des comportements et des consommations d'énergie finale.

- Groupe 1 : ce groupe comprend plutôt des ménages avec enfants, vivant en tant que propriétaire dans une maison individuelle, chauffée à l'électricité. En termes de pratiques domestiques, l'observation des valeurs moyennes des VS permet de dire que les ménages associés ont un équipement, bas une présence moyenne et des comportements de restriction. In fine, les consommations par personne en énergie finale sont faibles.
- Groupe 2 : dans ce groupe, on observe plutôt des ménages composés de personnes seules, plutôt âgée, souvent retraitées et plutôt pauvres. Les ménages occupent une maison individuelle, chauffée au gaz et ont des consommations par personne élevées.
- Groupe 3 : les ménages composant ce groupe sont des couples avec enfants, aux revenus élevés, vivant dans une maison individuelle chauffée au gaz ou à l'électricité. Ces ménages sont plutôt très équipés, et ont une consommation par personne plutôt faible.
- Groupe 4 : les ménages composant ce groupe sont plutôt des couples sans enfants, retraités, aux revenus moyens et vivant dans un logement individuel.
- Groupe 5 : ce groupe est caractérisé par le fait que les ménages le composant occupent un appartement, souvent en logement social. Les ménages sont plutôt plus jeunes et sont des couples avec enfants ou des familles monoparentales.

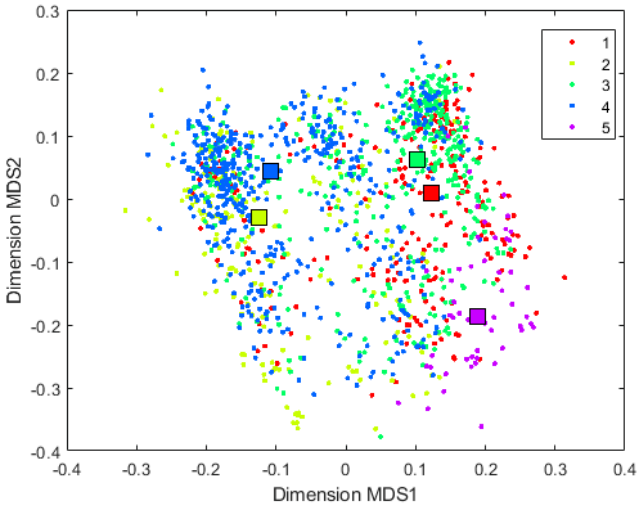


Figure 67 : représentation des situations d'habitation dans l'espace (MDS1, MDS2). Source : Auteur après calculs sur la base PHEBUS.

Le découpage décrit ici est intéressant dans le sens où il donne (logiquement) une place plus importante au nombre de personnes composant le ménage et donc la composition du ménage. Ce découpage n'occulte cependant pas les différences entre les logements individuels et collectifs et les types d'énergie utilisés pour le chauffage puisque ces caractéristiques servent également à discriminer les classes. Une dernière remarque sur les coefficients des modèles de régression peut être faite pour conclure cette étape de modélisation (Figure 68). On remarque que le coefficient associé à la variable surface est le même

pour tous les classes sauf pour ceux composés de ménages en fin de cycle de vie (classes 2 et 4) où le coefficient de surface est 50 à 100% plus important.

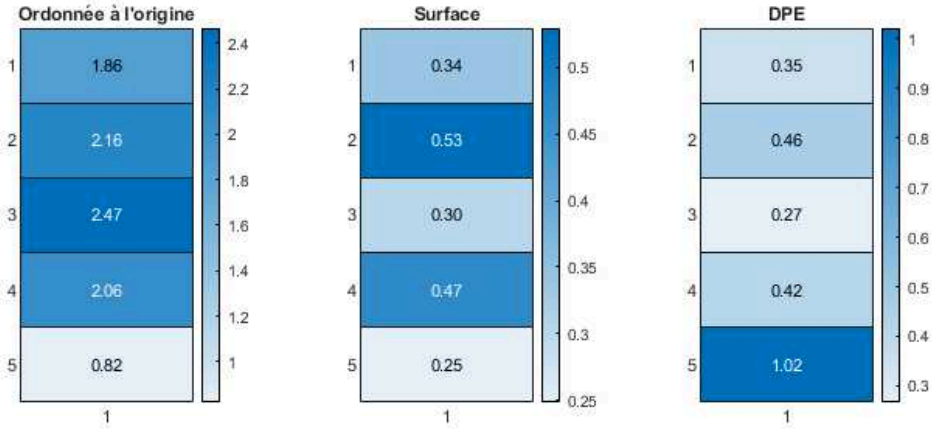


Figure 68 : Inventaire des coefficients de régression calculés pour le modèle de l'indicateur FECP. Les coefficient sont ordonnés par variable d'intérêt (colonne). Les types sont en lignes. Source : Auteur après calculs sur la base PHEBUS.

La modélisation réalisée ici complète la série des trois modèles MRHLP réalisée pour chacun des indicateurs de consommation. On retire ici comme résultat qu'il est possible en séparant les variables décrivant les situations d'habitation de construire un modèle de régression de bonne qualité à partir de seulement deux variables de régression caractérisant la surface et l'isolation thermique des logements. Ces deux variables apparaissent ainsi comme déterminantes pour prédire les consommations. Toutefois, nous avons observé que le partitionnement « optimal » (au sens du critère AIC dans le cadre du modèle MRHLP) variait selon l'indicateur considéré. Si les partitions fournies par les modèles des indicateurs FEC et FECM2 sont relativement proches (elles séparent les logements selon leur type et le type de chauffage), elles sont différentes du partitionnement calculé pour ce 3^{e} modèle (qui donne un primat à la composition du ménage).

Les partitions n'étant pas les mêmes, il pourrait être intéressant de les croiser pour construire un partitionnement plus fin permettant de décrire des situations d'habitation pour lesquelles les modèles des 3 indicateurs sont partagés. Cela pourrait être effectué en croisant les modèles calculés dans les parties précédentes. Toutefois, pour tester et étudier la performance de l'algorithme MRHLP nous choisissons de réaliser un modèle MRHLP « global » qui réalise simultanément le partitionnement des situations d'habitations et la régression des logarithmes des indicateurs FEC, FECP et FECM2.

3.8 Construction du modèle MRHLP des trois indicateurs de consommation

Sélection du modèle

La construction du modèle est effectuée de manière similaire. Les variables utilisées sont les suivantes :

- $\boldsymbol{X_1} = [\mathbf{1}, \boldsymbol{MDS}, \boldsymbol{MDS^2}]$
- $\boldsymbol{X_2} = [\mathbf{1}, \boldsymbol{log}(\boldsymbol{SURF}), \boldsymbol{log}(\boldsymbol{DPE})]$
- $\boldsymbol{Y} = [log(\boldsymbol{FEC}), log(\boldsymbol{FECM2}), log(\boldsymbol{FECP})]$

En effectuant 10 calculs successifs pour chaque valeur de K entre 2 et 14, on peut tracer l'évolution des critères d'information et d'estimation (Figure 69). Contrairement aux 3 modèles présents, les critères d'information (AIC et BIC) n'indiquent pas un minimum intéressant : l'ajout de clusters semble ajouter de l'information pertinente au regard de ces critères. Lorsqu'on observe les critères de précision pour chacun des indicateurs on observe en revanche que l'erreur quadratique pour l'indicateur FECP décroit significativement jusqu'à $K = 10$. L'analyse pour cette valeur de K donne de bons résultats, toutefois certains clusters sont redondants. On préfère garder finalement $K = 7$.

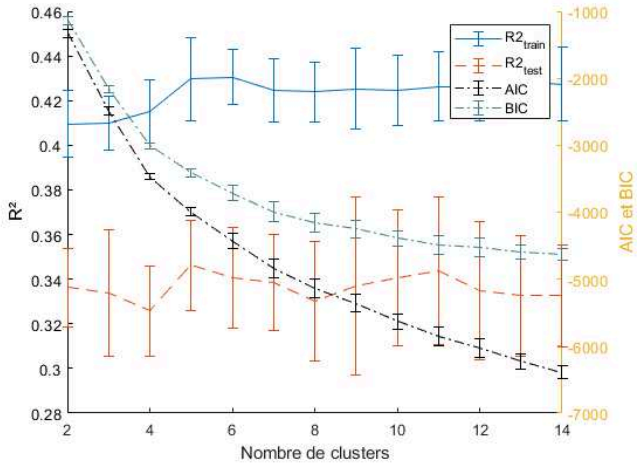


Figure 69 : Evolution des indicateurs d'information (AIC, BIC) et d'erreur (R²) en fonction du nombre de clusters, pour le modèle MRHLP modélisant les 3 indicateurs. Les moyennes et les écarts-types des valeurs sont représentées à partir de 10 calculs pour chaque valeurs de K. Source : Auteur après calculs sur la base PHEBUS.

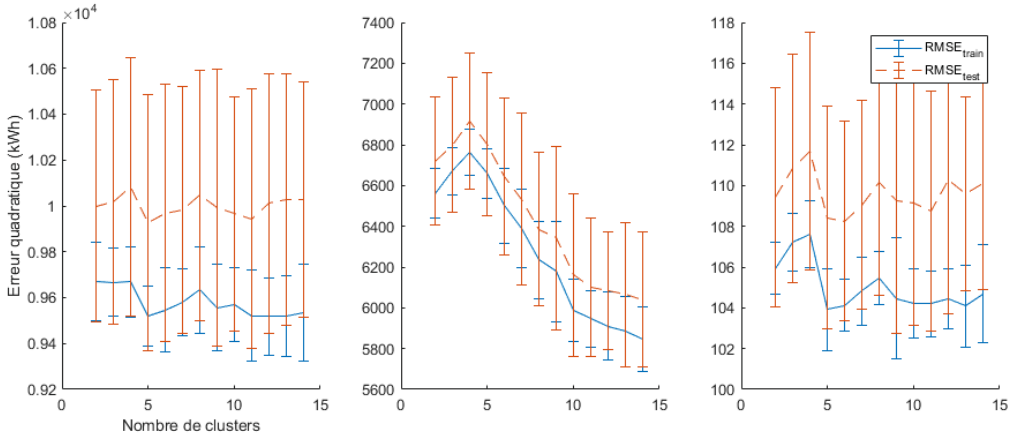


Figure 70 : Distribution des erreurs quadratiques pour les 3 indicateurs sur les sets d'entraînement et de test, par ordre de gauche à droite (RMSE pour FEC, FECP et FECM2) Source : Auteur après calculs sur la base PHEBUS.

En termes de performance d'estimation, le modèle offre des performances inférieures lorsqu'on le compare aux modèles de chacun des indicateurs. Cela peut être dû à l'addition des contraintes posées dans ce dernier modèle à savoir une contrainte sur les modèles liant les indicateurs aux paramètres explicatifs, une contrainte d'homogénéité des échantillons, une contrainte sur les niveaux des indicateurs, une contrainte posée par le fait de segmenter l'espace à partir des trois indicateurs simultanément. Les performances restent toutefois très convenables (comme en témoigne les indicateurs de performance des estimateurs à la Figure 71) et nous pouvons poursuivre l'analyse du modèle. Les améliorations potentielles sont discutées dans la fin de cette partie.

Répartition des lignes de la base d'entraînement

Cluster 1 : 124 lignes
Cluster 2 : 393 lignes
Cluster 3 : 205 lignes
Cluster 4 : 255 lignes
Cluster 5 : 135 lignes
Cluster 6 : 149 lignes
Cluster 7 : 127 lignes

Performances du modèle retenu

$AIC = -4682$

Indicateur	FEC	FECP	FECM2
$R^2{}_{train}$	0,43	0,32	0,32
$R^2{}_{test}$	0,36	0,26	0,3

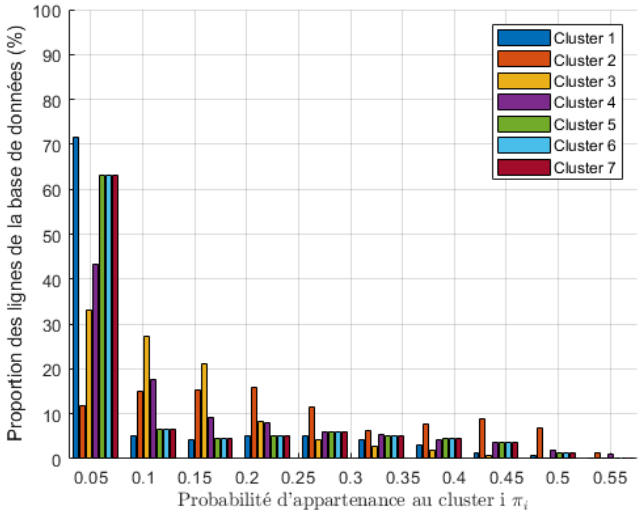


Figure 71 : Informations descriptives du modèle retenu. A gauche le nombre d'individus par type est donné. Les performances sur les données d'entrainement et de test sont données pour les trois indicateurs. A droite, on décrit la distribution des paramètres π_{ik}. Source : Auteur après calculs sur la base PHEBUS.

Analyse du modèle optimal

On présente le partitionnement et le modèle de régression obtenu pour $K = 7$. On observe sur la Figure 72 la distribution des ménages et des logements dans les différents types construits. Une observation générale peut être de distinguer 4 pôles sur cette « carte » :

- un pôle P1 regroupant les types 1 et 5
- un pôle P2 central regroupant les types 2 et 3
- un pôle P3 regroupant les types 4 et 7
- un pôle P4 constitué par le type 6

Cette division de l'espace MDS réalisée par le partitionnement peut être comparé à ce que l'on avait observé dans les 3 modèles précédents, c'est-à-dire une séparation des logements collectifs (P4) et des logements individuels (P1, P2 et P3). Puis une séparation en 3 pôles organisés selon la position dans le cycle de vie du ménage (P1 puis P2 et P3). On réalise dans la suite de ce paragraphe une analyse type par type des caractéristiques des logements, des ménages, des comportements et des consommations d'énergie. Pour chacun des types on observe la distribution des indicateurs de consommation au sein de chacun des types et on indique la position relative de cette distribution par rapport à la moyenne (élevée, basse). Une dénomination des types est proposée à partir de l'analyse des situations d'habitation et des pratiques domestiques.

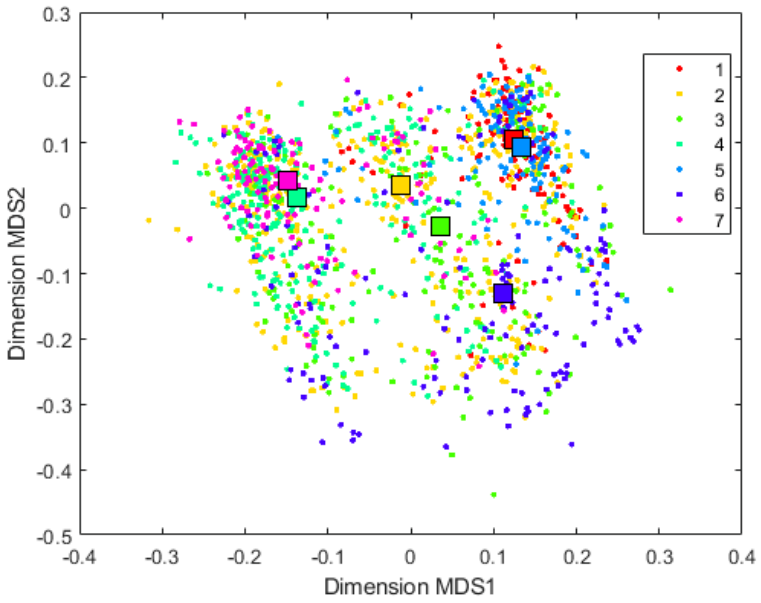


Figure 72 : Représentation des ménages et des logements dans l'espace MDS. Un code couleur permet de visualiser l'appartenance à un type. Source : Auteur après calculs sur la base PHEBUS.

- **Type 1 : Maison confortable**

Indicateurs	FEC élevée	FECM2 élevée	FECP basse

Ce groupe comprend essentiellement des couples d'actifs avec enfants et ayant des revenus élevés. Ces ménages occupent en tant que propriétaires une maison individuelle en zone urbaine ou rurale qui est chauffée au gaz. Ces ménages possèdent un grand nombre d'équipements et de bonne qualité. Ils ont une demande en chauffage moyenne, mais une consommation en ECS élevée. *In fine* ces ménages présentent des consommations totales très importantes mais des consommations par personne inférieures à la moyenne.

- **Type 2 : Maison froide**

Indicateurs	FEC basse	FECM2 basse	FECP basse

Ce groupe est constitué d'un diversité importante de situations d'habitations. Situé « au centre » de l'espace MDS avec le type 3, il est cependant caractérisé par le fait qu'il n'y a que des maisons individuelles. Les ménages composant ce groupe partagent aussi une faible demande en chauffage, des gestes de régulation et de restriction qui aboutissent à des consommations énergétiques globalement basses.

- **Type 3 : Appartement froid**

Indicateurs	FEC très basse	FECM2 très basse	FECP très basse

Ce groupe est d'une certaine manière le jumeau du type précédent à l'exception du fait qu'il comprend un nombre d'appartement très significatif. Au sein de ce cluster, les pratiques de régulation, voire de restriction amènent également à des consommations très faibles.

- **Type 4 : Cocon pratique**

Indicateurs	FEC élevée	FECM2 élevée	FECP élevée

Ce groupe comprend essentiellement des couples sans enfants ou des personnes seules, plutôt âgées et retraités et vivant comme propriétaires dans une maison individuelle de plus de 100 m² chauffée au gaz. Ces ménages plutôt pauvres ont tendanciellement un équipement moyen, une présence importante au logement, une demande en chaleur plutôt élevée. In fine, en associant une forte présence, un logement de grande taille et souvent mal isolé, ces ménages présentent des consommations très élevées.

- **Type 5 : Maison économe**

Indicateurs	FEC moyenne	FECM2 moyenne	FECP basse

Dans ce groupe on retrouve exclusivement des couples avec enfants aux revenus élevés et occupant une maison individuelle avec un DPE relativement bon (la médiane de ce type est de 205 kWh/m² à comparer à la médiane de l'échantillon total de 253 kWh/m²). Ainsi, malgré un nombre important de personnes,

un niveau d'équipement moyen le plus élevé entre tous les types, un niveau de régulation moyen, une présence moyenne au logement ces ménages ont une consommation énergétique moyenne.

- **Type 6 : Appartement chaud**

Indicateurs	FEC moyenne	FECM2 élevée	FECP moyenne

Au sein de ce type on observe une diversité importante de ménages occupant tous cependant un appartement. Ces ménages ont en général des revenus moyens mais partagent aussi le fait d'avoir une demande en chaleur relativement importante et peu de gestes de régulation. Ces ménage présentent in fine des consommations par mètre carré très élevées.

- **Type 7 : Cocon chaud**

Indicateurs	FEC moyenne	FECM2 moyenne	FECP élevée

Dans ce groupe, on trouve presque exclusivement des couples âgés, sans enfant au domicile, vivant dans une maison individuelle qu'ils possèdent. Ces ménages sont plutôt très équipés, sont très présents au logement et ont globalement peu de gestes de régulation. *In fine* ces ménages ont une consommation relativement élevée.

Ainsi, le calcul d'un modèle de régression faisant l'hypothèse de dynamiques hétérogènes (et latentes) permet d'identifier 7 groupes de situations d'habitations à partir de trois dimensions : le type de logement, le type de ménage et la demande en chaleur.

L'identification de ces classes peut être utile pour nourrir des travaux de recherche ou des réflexions visant à améliorer le ciblage des politiques du logement et de l'énergie. Ce modèle pourra servir aussi à la construction de travaux de prospective. Il pourrait être possible par exemple de réfléchir à l'évolution des types identifiés ainsi qu'à l'évolution des consommations associées (qui se traduirait par exemple par des hypothèses sur les évolutions des coefficients de régression). On donne dans le paragraphe suivant un résultat complémentaire issu des modèles calculés dans ce chapitre qui nourriraient ce type de réflexion.

3.9 Analyse complémentaire : les effets spatialisés de la surface sur la CED

Dans les paragraphes précédents nous avons centré l'analyse sur les types construits et l'identification des situations d'habitation. Il est cependant aussi intéressant de se replonger dans l'étude des modèles de régression construits. En particulier, les modèles MRHLP ont la particularité de calculer des coefficients pour chaque variable et pour chaque type. Par exemple, un modèle à 3 types aura 3 coefficients de surface. Le lecteur remarquera aussi que le calcul de la CED utilise une pondération des sous-modèles de régression pour fournir une estimation pour chaque situation d'habitation. Il est alors possible de calculer une « sensibilité » (un coefficient de régression) pour chaque individu en calculant le coefficient pondéré pour chaque variable. On obtient alors une spatialisation de « l'effet » d'une

variable. La formule mathématique permettant de calculer cet effet β_{eq} pour chaque situation d'habitation i est :

$$\beta_{eq}(i) \sum_{k=1}^{K} \pi_k(x_{1,i})\, \beta_k$$

On donne ci-dessous un exemple de représentation de ce coefficient pour la variable « surface ». On représente ce coefficient « spatialisé » pour les 3 indicateurs sur la Figure 73. Sur le plan méthodologique, en considérant que les modèles de régression étaient plus performants en termes de prédiction que le modèle global, nous avons choisi de représenter les graphes obtenus pour les modèles mono-variés. L'interprétation des coefficients peut être faite de la même manière que pour la régression multilinéaire. Le graphique (a) permet par exemple d'observer que l'effet du logarithme de la surface sur le logarithme décimal de la consommation d'énergie finale totale (FEC) varie entre 0,75 (pour les maison individuelles) et 1 (pour les logement collectifs). Ce chiffre est légèrement supérieur à ce qui a été trouvé dans la régression multilinéaire (respectivement 0,55 et 0,88). Selon ce graphique, cet effet n'est cependant pas discontinu et un ensemble large de situations d'habitation ont des coefficients intermédiaires ce qui suggère l'idée que les services énergétiques déployés dans les espaces domestiques (dit autrement, les intensités d'usage des équipements dans les logements) ne sont pas aussi nombreux partout.

Le croisement de cette observation avec les deux autres graphiques apporte un éclairage intéressant. Le graphique (b) montre une dynamique similaire en séparant les logements collectifs et les logements individuels. Il montre en particulier que l'élasticité de la consommation par mètre carré est plus faible pour les logements plus petits. Ce graphique ajoute cependant que cette élasticité négative (la consommation d'énergie par mètre carré diminue lorsque la surface augmente) est plus importante chez les personnes seules (graphique b). Cela est cohérent avec le fait qu'une personne seule ne développe pas autant de nouveaux services énergétiques lorsque la surface augmente que pourrait le faire un ménage plus nombreux.

Enfin l'analyse du graphique (c) permet de confirmer une dynamique différente de l'indicateur FECP avec la surface du logement selon la situation d'habitation considérée. En effet, on observe que la consommation en énergie finale par personne a une élasticité double chez les couples sans enfants et les personnes seules par rapport aux ménages avec enfants.

La surface joue donc un rôle différent sur la consommation d'énergie, selon la comptabilité énergétique que l'on considère (totale, par personne, par mètre carré) : l'analyse permet de confirmer le caractère structurant du lien entre le type de ménage et le logement dans la détermination de la CED. Sur le plan prospectif, la modélisation proposée ouvre selon nous des perspectives intéressantes de discussion sur les évolutions croisées de chacun des indicateurs pour les différentes situations d'habitation. En effet, utiliser simultanément les trois indicateurs et les mettre en regard des situations d'habitation permet de

nourrir et soutenir un ensemble d'hypothèses sur le futur des consommations d'énergie à partir des hypothèses de transformation (évolution de la surface, consécutivement à un déménagement, de l'isolation thermique etc.).

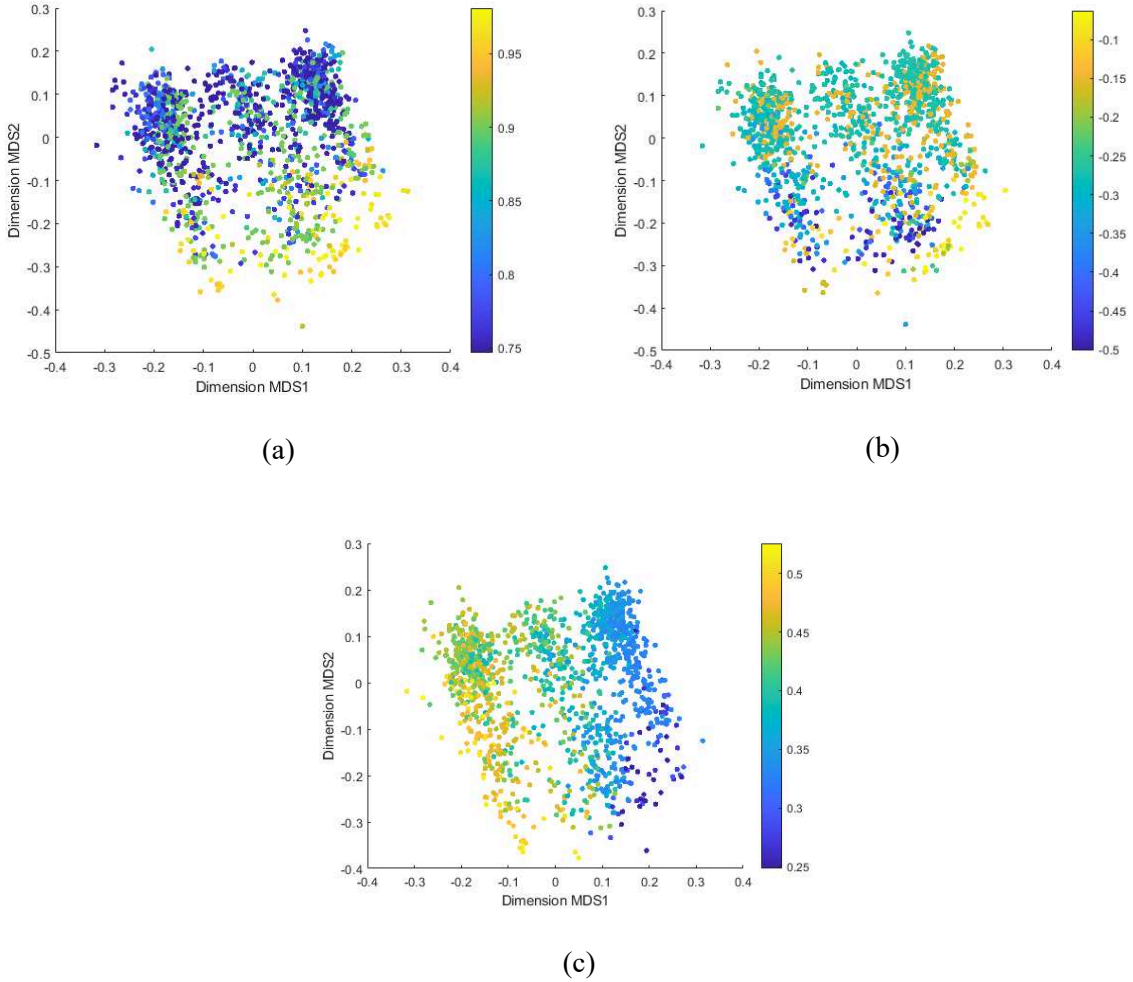


Figure 73 : Distribution du coefficient pour la variable surface et pour les trois indicateurs de consommation (FEC figure a, FECM2 figure b et FECP figure c). Une couleur chaude (resp. froide) indique une forte (resp. faible) sensibilité de l'indicateur de CED à la surface pour le ménage considéré. Source : Auteur après calculs sur la base PHEBUS.

4. Conclusion du chapitre

Dans ce chapitre, nous avons exploré la modélisation de la consommation d'énergie en articulant contextes résidentiels, pratiques énergétiques et consommations d'énergie. La construction d'un modèle s'est faite en 3 temps.

Premier temps : une modélisation linéaire de la CED pour explorer le set de données PHEBUS

En nous appuyant sur la littérature, nous avons d'abord construit des modèles linéaires pour explorer la base de données et observer les sélections des variables et analyser leurs effets. Nous avons pu d'une part observer le caractère prépondérant des variables caractérisant les contextes résidentiels puisque les

variables décrivant le type du logement, la composition du ménage, la surface, le type de chauffage et son âge, le revenu sont apparues comme des variables structurantes au sens statistique.

Ensuite, en réduisant le calcul des modèles à des sous échantillons contrôlés par le type de logement (maison/appartement) ou de chauffage (gaz/électricité), nous avons observé la variabilité des coefficients des modèles linéaires.

Deuxième temps : une modélisation non linéaire pour caractériser le pouvoir explicatif de la base de données PHEBUS

Dans un deuxième temps et avant d'aborder la construction d'un modèle basé sur les contextes résidentiels, nous avons souhaité construire un modèle de CED non linéaire avec les meilleures performances d'estimation. En utilisant le modèle XGBoost nous avons montré que seuls près de 45% de la variance de la consommation d'énergie finale totale était explicable à l'aide des variables contenues dans notre base de données, avec une erreur quadratique de près de 9MWh. Si la construction d'un modèle plus précis ne s'avérait donc pas possible, il nous a paru intéressant de poursuivre les investigations pour construire un modèle qui calcule automatiquement le partitionnement optimal de la base de données et la régression linéaire, tout en maintenant une bonne qualité d'estimation.

Troisième temps : une modélisation hiérarchique et probabiliste de 3 indicateurs de consommation

En nous basant sur les travaux du chapitre 2 qui ont montré l'articulation entre pratiques domestiques et contextes résidentiels, nous avons proposé de construire un modèle de CED qui permette le calcul simultané d'une régression et d'un partitionnement des contextes résidentiels. En modélisant successivement les 3 indicateurs à l'aide du modèle MRHLP nous montrons que chacun d'entre eux identifie différentes classes. L'indicateur FEC différencie à la fois le type de logement et la composition du ménage, tandis que les indicateurs FECP et FECM2 différencient respectivement plutôt le type de ménage et le type de logement. La modélisation croisée des 3 indicateurs a permis enfin d'identifier 7 classes de situations d'habitation sur lesquels une étude de la sensibilité à la surface a été réalisée.

Ce travail trouve de nombreuses limites et ouvre des perspectives de modélisation. En termes de limites, si ce travail a proposé des pistes de travail pour modéliser les pratiques énergétiques domestiques à l'aide de la méthode *ClustOfVar* et pour synthétiser un espace pour représenter les ménages et les logements à l'aide de la méthode MDS, l'étude de la sensibilité sur les résultats de ces choix méthodologiques serait à mener. Par ailleurs, il nous paraitrait intéressant de poursuivre le développement du modèle MRHLP en intégrant le calcul de l'espace et des variables synthétiques. Une seconde limite est liée à la qualité des données utilisées. Si nous avons fait l'effort de présenter les données, mettre en regard nos résultats avec ceux de la littérature, et étudier la qualité des données à travers les travaux connus, il faut concéder que nous avons peu de recul sur la qualité des données en

particulier du DPE. Par ailleurs, les données caractérisant les comportements domestiques pourraient être améliorées pour traduire au mieux les comportements réels.

En synthèse, ce travail de modélisation aura permis de voir qu'il est possible de construire un modèle centré sur les contextes résidentiels. Si le travail ne permet pas d'améliorer significativement la qualité de l'estimation du modèle, il permet de renforcer l'explication qui permet de comprendre le niveau des indicateurs de consommation, avec une performance d'estimation similaire. En particulier, cette modélisation quantitative des pratiques propose de mettre en interface les travaux quantitatifs menés dans le domaine de l'ingénierie et de l'économie avec les travaux en sociologie de l'énergie sur les pratiques énergétiques domestiques. Nous croyons qu'en proposant une typologie de situations d'habitation et en expliquant les niveaux de consommations selon 3 indicateurs en mettant en relation les ménages, les logements, les pratiques, le modèle ouvre une perspective intéressante de collaboration interdisciplinaire.

Conclusion générale

Le cadre de la thèse

Ce travail de thèse a été entrepris pour apporter des éléments de compréhension sur la consommation d'énergie dans l'espace domestique. En faisant le constat d'une pluralité d'approches disciplinaires et en s'appuyant sur les littératures en sociologie de l'énergie et de modélisation statistique, il apporte par la modélisation des éléments méthodologiques et des résultats heuristiques permettant d'améliorer notre compréhension des pratiques sociales et leurs liens avec la consommation d'énergie dans les logements.

La problématique posée dans ce travail a été déclinée en verrous scientifiques. En premier lieu, nous nous sommes interrogés (Q1) sur la façon dont les pratiques domestiques étaient liées aux situations d'habitations (les relations ménage-logement) et entre elles ? En second lieu, nous avons questionné (Q2) comment ces pratiques et leurs liens éventuels au sein de styles de vie résidentiels pouvaient être intégrées à des modèles de consommation d'énergie, pour estimer correctement les consommations ? Ces questions ont impliqué dans ce travail de recherche une réflexion et une contribution, en troisième lieu (Q3), sur l'intégration des pratiques sociales, souvent décrites à travers des variables qualitatives, dans un modèle quantitatif de consommation d'énergie.

Le travail de thèse a été organisé en trois chapitres : un premier temps de revue de littérature a permis de construire un cadre de modélisation centré sur les situations d'habitation et les pratiques des ménages. Il a été suivi d'un temps de modélisation permettant de mettre à l'épreuve ce cadre, en particulier sur le lien entre les pratiques, les styles de vie résidentiels, les situations d'habitation et les caractéristiques des ménages et des logements. Enfin, le troisième chapitre a présenté une proposition de modélisation quantitative de la consommation d'énergie domestique valorisant dans sa construction les résultats du deuxième chapitre.

Les résultats

Une revue de littérature interdisciplinaire qui met en avant le besoin de travaux de modélisation des pratiques sociales et de leurs liens avec les consommations énergétiques

La revue de littérature présentée dans le premier chapitre a mis en évidence un foisonnement d'approches issus de l'économie, de la psychologie, de la sociologie, des sciences de l'ingénieur, des sciences de la donnée. En revanche, le croisement de ces dernières a mis en évidences deux difficultés. Tout d'abord, le travail a montré le manque de travaux quantitatifs valorisant les concepts construits en sociologie de l'énergie, qui nourrissent des approches originales pour améliorer notre compréhension afin d'aider les décideurs publics. Il s'agit par exemple du parcours résidentiel du ménage, de la définition du confort comme une variable plurielle, latente et dynamique, et de l'influence des contextes résidentiels sur les pratiques. La revue de littérature a ensuite souligné la difficulté pour construire un

cadre de modélisation testé empiriquement et associant situations d'habitation, pratiques des ménages, consommations d'énergie.

La formulation d'un cadre de modélisation de la consommation d'énergie articulant un « système énergétique domestique » avec des « situations d'habitation »

Dans la synthèse de la revue de littérature nous nous sommes appuyés sur les travaux de Shove (2003) et Kowsari (2011) pour proposer un cadre de modélisation de la consommation énergétique. A l'échelle du ménage et du logement, le modèle propose de voir un « système énergétique domestique » qui est un ensemble cohérent constitué d'un environnement matériel (équipements, bâti), de pratiques (comportements, activités) et d'un environnement cognitif (savoir, normes, compétences). Ce système est au moins partiellement observable (à l'aide d'un travail d'enquête par exemple) à travers ces trois dimensions. A l'échelle macroscopique (d'une région par exemple), il est possible de considérer un ensemble de « systèmes énergétiques domestiques ». Dans notre modèle nous avons fait l'hypothèse que ces derniers étaient déterminés par les caractéristiques des individus, des ménages et des contextes (résidentiels, culturels, légaux).

Nous avons ensuite dû simplifier ce cadre de modélisation afin de pouvoir le valider avec les données d'enquête disponibles. Nous avons dû restreindre l'étude des systèmes énergétiques domestiques à celles des pratiques et des équipements et écarter la dimension cognitive. Un travail de validation a été mené en classifiant ces données pour des ménages vivant en Île de France (Enquête ENERGIHAB de 2010).

Une identification des « dimensions comportementales » qui fournit des éléments de compréhension et des éléments méthodologiques.

Le travail de modélisation mené au chapitre 2 a permis de discuter de la faisabilité technique et des résultats de la modélisation des pratiques sociales à partir de données d'enquête. Une classification des variables par analyse factorielle (de données mixtes) a été comparée avec une classification de variables selon un critère de corrélation. La comparaison a montré que dans ce cas la première méthode offre une meilleure stabilité selon l'échantillon que la seconde, mais présente des axes factoriels moins nombreux et moins facilement interprétables. Sur le plan méthodologique, le travail mené invite donc de manière générale à considérer les deux approches afin de comparer les résultats. Dans une approche de modélisation explicative, la seconde approche nous parait cependant plus pertinente en raison d'une plus forte portée explicative. Sur le plan de l'interprétation, l'analyse montre que les dimensions comportementales identifiées sont les niveaux : d'équipement dans le domaine de l'alimentation, d'occupation du logement, d'occupation moyenne du logement, d'équipement et d'usage des équipements d'hygiène ; et la demande en chauffage et son niveau de régulation, l'importance des comportements de restriction de chauffage, les gestes de régulation « verts », le type d'éclairage.

Une identification de styles de vie résidentiels et de situations d'habitation

Le travail de classification des données caractérisant les pratiques des ménages a permis de dégager quatre styles de vie résidentiels liés à des situations d'habitation spécifiques. Les styles de vie décrivent des modes de vie différenciés par les pratiques de régulation, le niveau d'équipement et l'occupation du logement. Les situations d'habitation identifiées sont caractérisées quant à elles selon le type de logement, le parcours résidentiel du ménage, sa composition et son revenu. Le travail de classification a montré d'une part une interrelation des variables de comportements entre elles et avec les situations d'habitation. Cette liaison n'est cependant pas suffisamment forte pour pouvoir faire une prévision exacte du style de vie résidentiel d'une situation d'habitation donnée. Des investigations complémentaires sur ce point sont proposées à la fin de cette conclusion.

Une modélisation hiérarchique de la consommation d'énergie domestique pour valider le cadre de modélisation et proposer une méthodologie originale de classification et de régression de données mixtes

Dans la dernière partie, les données de l'enquête PHEBUS collectées auprès de ménages français en 2012 ont été utilisées pour construire un modèle de la consommation d'énergie domestique. Le lien statistique entre les pratiques domestiques et les situations d'habitation étudié au chapitre 2 a été supposé vérifié à l'échelle nationale et une modélisation hiérarchique a été proposée. Dans un premier temps, nous avons détaillé la construction d'un modèle de régression multilinéaire comme modèle de référence. L'utilisation de ce dernier sur plusieurs sous-ensemble de données a montré que les performances d'estimation (R^2, RMSE) étaient très différentes selon que l'on considère des logements individuels ($\overline{R^2}_{maisons} = 45\%$) ou des logements collectifs ($\overline{R^2}_{appartements} = 59\%$). Par ailleurs, il apparaît que les coefficients de ces modèles, modélisant l'effet « toutes choses égales par ailleurs » des variables sont très différents notamment pour la surface et la variable modélisant la qualité de l'isolation thermique (ici le DPE). Nous pouvons faire l'hypothèse que la performance d'estimation de ce type de modèle sur un ensemble de données contenant à la fois des logement collectifs et individuels dépend de la proportion de ces types de logements dans l'échantillon. Ces calculs ont conforté les résultats du chapitre 2 qui invitaient à différencier les situations d'habitation pour étudier les consommations d'énergie. D'un autre côté, l'étude de la modélisation de la consommation d'énergie à travers d'autres indicateurs issus de la littérature comme l'énergie divisée par le nombre de personnes du ménage et par m² de surface du logement offrent des performances d'estimation et des coefficients très différents. Enfin, une modélisation de la consommation en énergie finale totale à l'aide d'un modèle d'apprentissage avancé (XGBoost) nous a permis de fournir une estimation des performances d'estimations « maximales ». Le modèle montre que les variables explicatives mobilisées permettent d'expliquer près de 45% de la variance de la variable expliquée.

Dans un second temps, une modélisation « intégrée » des consommations d'énergie est proposée. Celle-ci a été construite de manière à respecter deux contraintes : (1) le calcul simultané d'une segmentation des situations d'habitation et d'une régression de la consommation d'énergie, (2) la modélisation croisée des trois indicateurs de consommations d'énergie (totale, par personne et par m² du logement). La structure de modélisation choisie associe des régressions multilinéaires à proportions logistiques. Un premier modèle de régression linéaire est utilisé et deux variables sont mobilisées (surface et niveau d'isolation thermique). Liées statistiquement aux situations d'habitation, les variables décrivant les pratiques sont utilisées comme variables supplémentaires pour décrire les classes et les modèles de régression obtenus.

L'analyse du modèle construit permet de dégager plusieurs résultats. Tout d'abord le modèle identifie effectivement des modèles de régression différents sur des sous-ensembles de situations d'habitation. Ensuite, il apparaît que l'utilisation de différentes métriques permet d'identifier des classes différentes : la comptabilité en énergie finale totale différencie à la fois le type de logement et la composition du ménage, tandis que les indicateurs par personnes et par m² différencient respectivement surtout le type de ménage et le type de logement. *In fine*, le modèle final permet d'identifier sept situations d'habitations différenciées par le type de logement, le type de ménage et ses comportements de régulations. L'analyse des sept classes permet de formuler des hypothèses pour expliquer le niveau des indicateurs de consommation (comportements, équipements et/ou isolation et/ou taille du logement).

En termes d'analyse énergétique, l'homogénéité des classes de situations d'habitation, de pratiques et de niveaux de consommations plaide en faveur du cadre de modélisation proposé dans cette thèse. Remis dans le cadre d'une politique visant à diminuer les consommations d'énergie, les résultats soulignent l'importance d'une transformation des styles de vie résidentiels et des situations d'habitation.

Sur le plan méthodologique, le modèle présenté se base sur un algorithme initialement développé pour la segmentation de séries temporelles. Les hypothèses posées (notamment pour la construction d'un espace synthétique des situations d'habitation) devront être discutées mais le modèle propose une contribution pour améliorer les travaux de classification et les modèles de régression à partir de données (mixtes) d'enquête.

Perspectives

Ce travail ouvre de nombreuses perspectives de recherches que nous présentons synthétiquement.

Perspective n°1 : **Mettre à jour les résultats avec des données plus récentes.**

Une première discussion qui nous paraît importante à l'issue de ce travail de thèse est de mener une discussion sur la sensibilité des styles de vie résidentiels aux variables de comportements sélectionnées. Dans notre recherche, nous avons mobilisé les données d'enquêtes disponibles et sélectionnées par les experts de ce champ. Toutefois, il nous paraît important de les actualiser pour suivre les évolutions des

pratiques domestiques avec le contexte économique et social (développement du télétravail, expansion du numérique, renchérissement du prix de l'énergie, politiques publiques visant la « sobriété » énergétique etc.) et des évolutions des modes d'habiter (structure des ménages, qualité des logements, etc.). Ainsi, une réflexion pourrait être menée afin de mieux collecter dans les enquêtes auprès des ménages des données permettant de décrire le confort (la norme latente) des habitants. Cette réflexion pourrait conduire une collecte des données sur les matériaux, les couleurs utilisées dans les logements, les activités menées au sein du logement, etc.

Perspective n°2 : **Quantifier la sensibilité des résultats à la qualité des données et au modèle utilisé**

Dans une perspective de transfert de la méthodologie à d'autres équipes de recherche il serait intéressant de connaître quelle est l'influence de l'absence de chacune des variables dans le *set* de données sur les résultats. Si nous savons par expérience qu'elle n'est pas majeure dans notre cas, une étude de la robustesse serait intéressante pour l'état de l'art et guider les chercheurs et les opérateurs souhaitant réaliser ce travail de modélisation. De plus, la modélisation effectuée ici avec des données françaises pourrait être répliquée avec des données issues d'autres pays afin d'étudier les divergences. La difficulté résidera dans l'identification des meilleures variables d'enquête. Enfin, nous avons comparé ici plusieurs approches de modélisation mais nous n'avons pas exploré les nombreuses alternatives, dont les réseaux de neurones, qui pourraient vraisemblablement fournir des résultats intéressants.

Perspective n°3 : **Elargir le travail de modélisation à d'autres types de données**

Dans ce travail, nous avons valorisé essentiellement des données d'enquête par questionnaire. Il pourrait être intéressant d'utiliser des données issues d'autres source pour corroborer les analyses menées. Par exemple, croiser les styles de vie résidentiels avec des données issues des enquêtes Emploi du Temps réalisées périodiquement dans tous les pays par les instituts statistiques. Ainsi dans le cadre de projets plus restreints, il pourrait être intéressant d'utiliser des données de consommation mesurées pour d'une part avoir une meilleure connaissance des erreurs des données de consommation, et d'autre part pour alimenter le travail de modélisation.

Perspective n°4 : **Utiliser le cadre de la modélisation basé sur les « situations d'habitation » pour caractériser les effets de transformations des situations d'habitation (changement de comportement, isolation, adaptation du logement)**

Une autre valorisation importante pourrait consister à approfondir l'analyse du modèle en caractérisant l'effet de plusieurs variables (surface, isolation, comportements par exemple) au sein de chaque cluster. En effet, l'étude des effets restreint à ces sous-ensembles homogènes pourraient permettre de caractériser le degré d'influence des différentes stratégies (changement de comportement, réalisation de travaux d'isolation, adaptation du logement aux besoins du ménage). Cette étude pourrait par exemple être réalisée à l'aide des données de l'enquête ENL.

Perspective n°5 : **Utiliser le cadre de la modélisation basé sur les « situations d'habitation » pour étudier les dynamiques de transformation des styles de vie résidentiels**

Une limite importante de notre travail est que les données à notre disposition ne renseignent pas sur l'évolution des comportements énergétiques des ménages. Une enquête longitudinale qui permettrait de suivre simultanément les pratiques, les structures des situations d'habitation, les individus et les consommations permettrait de caractériser bien plus finement les dynamiques liées au cycle de vie du ménage et au changement du contexte résidentiel.

Perspective n°6 : **Permettre une évaluation énergétique des politiques publiques du logement et alimenter une réflexion prospective sur la consommation d'énergie du secteur résidentiel**

Le cadre de modélisation proposé dans cette thèse peut être valorisé à travers des travaux prospectifs. Ceux-ci se nourrissent en effet de briques de modélisation disciplinaires pour fournir une évaluation quantifiée de scénarios énergétiques. En particulier, l'originalité de ce modèle est qu'il permet un croisement avec des modèles de peuplement. Aussi, il serait intéressant de valoriser les données de l'enquête Logement pour quantifier et caractériser rétrospectivement les effets des différentes politiques du logement sur les situations d'habitation et les consommations énergétiques à travers le prisme proposé dans cette thèse.

La modélisation à l'aide des données de l'Enquête Logement permettrait par ailleurs d'alimenter la discussion autour de problématiques clés dans le secteur du logement comme « l'effet rebond » et l'« energy efficiency gap ».

Perspective n°7 : **Contribuer au développement et l'amélioration des modèles et algorithmes utilisant des données mixtes.**

Ce travail de thèse s'est attaché à manipuler des données mixtes afin de modéliser les pratiques sociales. Plusieurs modèles ont été utilisés et comparés et nous avons pu voir des perspectives originales de modélisation. En particulier, le modèle de Selosse (2020) utilisé dans le chapitre 2 nous parait très prometteur, même si les résultats présentés dans ce travail sont plus difficilement exploitables et éloignés des résultats des autres méthodes. La formulation probabiliste du modèle de co-clustering pourrait être enrichie en construisant simultanément une variable synthétique pour chaque groupe de variables. Cette variable synthétique permettrait de dériver les variables rattachées à son groupe à l'aide de lois de probabilité paramétrées (et dont la nature serait liée à la nature des variables originales).

Liste des figures

Liste des tableaux

Bibliographie

Aerts, D., J. Minnen, I. Glorieux, I. Wouters, and F. Descamps. 2014. 'A Method for the Identification and Modelling of Realistic Domestic Occupancy Sequences for Building Energy Demand Simulations and Peer Comparison'. *Building and Environment* 75 (May): 67–78. https://doi.org/10.1016/j.buildenv.2014.01.021.

Ahmad, A. S., M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin, H. Abdullah, and R. Saidur. 2014. 'A Review on Applications of ANN and SVM for Building Electrical Energy Consumption Forecasting'. *Renewable and Sustainable Energy Reviews* 33 (May): 102–9. https://doi.org/10.1016/j.rser.2014.01.069.

Ahmed Gassar, Abdo Abdullah, Geun Young Yun, and Sumin Kim. 2019. 'Data-Driven Approach to Prediction of Residential Energy Consumption at Urban Scales in London'. *Energy* 187 (C). https://ideas.repec.org/a/eee/energy/v187y2019ics0360544219316639.html.

Aigner, Dennis J., Cynts Sorooshian, and Pamela Kerwin. 1984. 'Conditional Demand Analysis for Estimating Residential End-Use Load Profiles'. *The Energy Journal* 5 (3). https://doi.org/10.5547/ISSN0195-6574-EJ-Vol5-No3-6.

Ajzen, Icek. 1985. 'From Intentions to Actions: A Theory of Planned Behavior'. In *Action Control: From Cognition to Behavior*, edited by Julius Kuhl and Jürgen Beckmann, 11–39. SSSP Springer Series in Social Psychology. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-69746-3_2.

Alberini, Anna, Will Gans, and Daniel Velez-Lopez. 2011. 'Residential Consumption of Gas and Electricity in the U.S.: The Role of Prices and Income'. *Energy Economics* 33 (5): 870–81. https://doi.org/10.1016/j.eneco.2011.01.015.

Allibe, Benoit. 2012. 'Modélisation des consommations d'énergie du secteur résidentiel français à long terme - Amélioration du réalisme comportemental et scénarios volontaristes', Thèse de doctorat. Ecole des Hautes Etudes en Sciences Sociales (EHESS).

Amasyali, Kadir, and Nora M. El-Gohary. 2018. 'A Review of Data-Driven Building Energy Consumption Prediction Studies'. *Renewable and Sustainable Energy Reviews* 81 (January): 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

Andersen, Rune Korsholm. 2012. 'The Influence of Occupants' Behaviour on Energy Consumption Investigated in 290 Identical Dwellings and in 35 Apartments'. Abstract from 10th International Conference on Healthy Buildings. https://www.researchgate.net/publication/255709305.

Aoun, Nadine, Roland Bavière, Mathieu Vallée, Adrien Brun, and Guillaume Sandou. 2019. 'Dynamic Simulation of Residential Buildings Supporting the Development of Flexible Control in District Heating Systems'. In Linköping Electronic Conference Proceedings 157:13, p. 10, 129–38. https://doi.org/10.3384/ecp19157129.

Aune, Margrethe. 2007. 'Energy Comes Home'. *Energy Policy* 35 (11): 5457–65. https://doi.org/10.1016/j.enpol.2007.05.007.

Bachelard, Gaston. 1934. *Le Nouvel Esprit Scientifique*. Paris: PUF.

Ballarini, Ilaria, Stefano Paolo Corgnati, and Vincenzo Corrado. 2014. 'Use of Reference Buildings to Assess the Energy Saving Potentials of the Residential Building Stock: The Experience of TABULA Project'. *Energy Policy* 68 (May): 273–84. https://doi.org/10.1016/j.enpol.2014.01.027.

Belaïd, Fateh. 2017. 'Untangling the Complexity of the Direct and Indirect Determinants of the Residential Energy Consumption in France: Quantitative Analysis Using a Structural Equation

Modeling Approach'. *Energy Policy* 110 (November): 246–56. https://doi.org/10.1016/j.enpol.2017.08.027.

Ben, Hui, and Koen Steemers. 2018. 'Household Archetypes and Behavioural Patterns in UK Domestic Energy Use'. *Energy Efficiency* 11 (3): 761–71. https://doi.org/10.1007/s12053-017-9609-1.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert. 2000. 'Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (7): 719–25. https://doi.org/10.1109/34.865189.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York, NY: Springer.

Bogin, Diana, Meidad Kissinger, and Evyatar Erell. 2021. 'Comparison of Domestic Lifestyle Energy Consumption Clustering Approaches'. *Energy and Buildings* 253 (December): 111537. https://doi.org/10.1016/j.enbuild.2021.111537.

Bonnin, Marguerite. 2016. 'Habitable et Confortable : Modèles Culturels, Pratiques de l'habitat et Pratiques de Consommation d'énergie En Logement Social et Copropriétés'. These de doctorat, Paris 10. https://www.theses.fr/2016PA100003.

Bourgeois, Alexis, Margot Pellegrino, and Jean-Pierre Lévy. 2017. 'Modeling and Mapping Domestic Energy Behavior: Insights from a Consumer Survey in France'. *Energy Research & Social Science* 32 (October): 180–92. https://doi.org/10.1016/j.erss.2017.06.021.

Bovay, Claude. 1987. *L'Energie au quotidien: aspects sociologiques et éthiques de la consommation d'énergie*. Labor et Fides.

Buttitta, Giuseppina, William J. N. Turner, Olivier Neu, and Donal P. Finn. 2019. 'Development of Occupancy-Integrated Archetypes: Use of Data Mining Clustering Techniques to Embed Occupant Behaviour Profiles in Archetypes'. *Energy and Buildings* 198 (September): 84–99. https://doi.org/10.1016/j.enbuild.2019.05.056.

Cavailhès, Jean, Daniel Joly, Thierry Brossard, Hervé Cardot, Mohammed Hilal, and Pierre Wavresky. 2011. 'La Consommation d'énergie Des Ménages En France'. https://www.precarite-energie.org/IMG/pdf/Rapport_final.pdf.

Certeau, Michel De, Fredric Jameson, and Carl Lovitt. 1980. 'On the Oppositional Practices of Everyday Life'. *Social Text*, no. 3: 3. https://doi.org/10.2307/466341.

Chamroukhi, F., H. Glotin, and A. Samé. 2013. 'Model-Based Functional Mixture Discriminant Analysis with Hidden Process Regression for Curve Classification'. *Neurocomputing*, Advances in artificial neural networks, machine learning, and computational intelligence, 112 (July): 153–63. https://doi.org/10.1016/j.neucom.2012.10.030.

Chavent, Marie, V. Kuentz Simonet, Benoit Liquet, and Jérôme Saracco. 2012. 'ClustOfVar: An R Package for the Clustering of Variables'. *Journal of Statistical Software* 50 (13): 1–16.

Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2023. 'Xgboost: Extreme Gradient Boosting'. https://cran.r-project.org/web/packages/xgboost/index.html.

Chiu, Lai Fong, Robert Lowe, Rokia Raslan, Hector Altamirano-Medina, and Jez Wingfield. 2014. 'A Socio-Technical Approach to Post-Occupancy Evaluation: Interactive Adaptability in Domestic Retrofit'. *Building Research & Information* 42 (5): 574–90. https://doi.org/10.1080/09613218.2014.912539.

Chou, Jui-Sheng, and Dac-Khuong Bui. 2014. ‘Modeling Heating and Cooling Loads by Artificial Intelligence for Energy-Efficient Building Design’. *Energy and Buildings* 82 (October): 437–46. https://doi.org/10.1016/j.enbuild.2014.07.036.

Christiansen, Gerhard, Dr Manuel Frondel, Peter Grösche, Dr Harald Tauchmann, Dr Colin Vance, and Ute Müller. 2005. ‘The German Residential Energy Consumption Survey 2005’. Research Projec 15/06. RWI Essen: Rheinisch-Westfälisches Institut für Wirtschaftsforschung.

Cowan, Ruth Schwartz. 1985. *More Work For Mother: The Ironies Of Household Technology From The Open Hearth To The Microwave*. Basic Books.

De Soete, Geert, and J. Douglas Carroll. 1994. ‘K-Means Clustering in a Low-Dimensional Euclidean Space’. In *New Approaches in Classification and Data Analysis*, edited by Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand, and Bernard Burtschy, 212–19. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-51175-2_24.

Delahais, Thomas, and Agathe Devaux-Spatarakis. 2018. ‘Évaluation des politiques publiques et sociologie : état des lieux d’une relation distanciée’. *Sociologies pratiques* 36 (1): 47–56. https://doi.org/10.3917/sopr.036.0047.

Denis, C, and F Varenne. 2019. ‘Interprétabilité et explicabilité pour l’apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique.’ In *Actes de la CNIA PFIA*, 9. Toulouse: AFIA.

Desrosières, Alain. 1995. ‘Classer et mesurer : les deux faces de l’argument statistique’. *Réseaux. Communication - Technologie - Société* 13 (71): 11–29. https://doi.org/10.3406/reso.1995.2689.

———. 2001. ‘Entre réalisme métrologique et conventions d’équivalence : les ambiguïtés de la sociologie quantitative’. *Genèses* 43 (2): 112–27. https://doi.org/10.3917/gen.043.0112.

———. 2013. *Pour une sociologie historique de la quantification: L’Argument statistique I*. Presses des Mines via OpenEdition.

Diao, Longquan, Yongjun Sun, Zejun Chen, and Jiayu Chen. 2017. ‘Modeling Energy Consumption in Residential Buildings: A Bottom-up Analysis Based on Occupant Behavior Pattern Clustering and Stochastic Simulation’. *Energy and Buildings* 147 (July): 47–66. https://doi.org/10.1016/j.enbuild.2017.04.072.

Dubuisson-Quellier, Sophie, and Marie Plessz. 2013. ‘La théorie des pratiques. Quels apports pour l’étude sociologique de la consommation ?’ *Sociologie*, no. N°4, vol. 4 (December). http://journals.openedition.org/sociologie/2030.

Eon, Christine, Jessica Breadsell, Gregory Morrison, and Joshua Byrne. 2019. ‘Shifting Home Energy Consumption Through a Holistic Understanding of the Home System of Practice’. In *Decarbonising the Built Environment: Charting the Transition*, edited by Peter Newton, Deo Prasad, Alistair Sproul, and Stephen White, 431–47. Singapore: Springer. https://doi.org/10.1007/978-981-13-7940-6_23.

Estiri, Hossein. 2015. ‘A Structural Equation Model of Energy Consumption in the United States: Untangling the Complexity of per-Capita Residential Energy Use’. *Energy Research & Social Science* 6 (March): 109–20. https://doi.org/10.1016/j.erss.2015.01.002.

Frederiks, Elisha R., Karen Stenner, and Elizabeth V. Hobman. 2015. ‘The Socio-Demographic and Psychological Predictors of Residential Energy Consumption: A Comprehensive Review’. *Energies* 8 (1): 573–609. https://doi.org/10.3390/en8010573.

Friedman, Jerome H. 1991. ‘Multivariate Adaptive Regression Splines’. *The Annals of Statistics* 19 (1): 1–67. https://doi.org/10.1214/aos/1176347963.

Gaetani, Isabella, Pieter-Jan Hoes, and Jan L. M. Hensen. 2016. 'Occupant Behavior in Building Energy Simulation: Towards a Fit-for-Purpose Modeling Strategy'. *Energy and Buildings* 121 (June): 188–204. https://doi.org/10.1016/j.enbuild.2016.03.038.

Garabuau-Moussaoui, Isabelle. 2009. 'Behaviours, Transmissions, Generations: Why Is Energy Efficiency Not Enough?' 1 In *ECEEE Summer Study Proceedings* (pp. 33-43). European Council for an Energy-Efficient Economy, Stockholm, Sweden.

Geneviève, Lacroix. 2004. 'Analyse Conditionnelle de La Demande Appliquée Au Secteur Résidentiel Québécois En 1989, 1994, et 1999'. Unviersité de Laval. https://www.collectionscanada.gc.ca/obj/s4/f2/dsk4/etd/MQ98138.PDF?oclc_number=63124763.

Giraudet, Louis-Gaëtan, Céline Guivarch, and Philippe Quirion. 2012. 'Exploring the Potential for Energy Conservation in French Households through Hybrid Modeling'. *Energy Economics* 34 (2): 426–45. https://doi.org/10.1016/j.eneco.2011.07.010.

Glotin, David, Cyril Bourgeois, Louis-Gaëtan Giraudet, and Philippe Quirion. 2019. 'Prediction Is Difficult, Even When It's about the Past: A Hindcast Experiment Using Res-IRF, an Integrated Energy-Economy Model'. *Energy Economics*, Eighth Atlantic Workshop on Energy and Environmental Economics, 84 (October): 104452. https://doi.org/10.1016/j.eneco.2019.07.012.

Gohier, Christiane. 2004. 'De la démarcation entre critères d'ordre scientifique et d'ordre éthique en recherche interprétative'. *Recherches qualitatives* 24: 3–17. https://doi.org/10.7202/1085561ar.

Gouvernement. 2022. 'Plan de Sobriété Énergétique - Dossier de Presse'. https://www.ecologie.gouv.fr/sites/default/files/dp-plan-sobriete.pdf.

Gower, J. C. 1971. 'A General Coefficient of Similarity and Some of Its Properties'. *Biometrics* 27 (4): 857–71. https://doi.org/10.2307/2528823.

Guerra Santin, Olivia. 2011. 'Behavioural Patterns and User Profiles Related to Energy Consumption for Heating'. *Energy and Buildings* 43 (10): 2662–72. https://doi.org/10.1016/j.enbuild.2011.06.024.

Hache, Emmanuel, Déborah Leboullenger, and Valérie Mignon. 2017. 'Beyond Average Energy Consumption in the French Residential Housing Market: A Household Classification Approach'. *Energy Policy* 107 (August): 82–95. https://doi.org/10.1016/j.enpol.2017.04.038.

Hackett, Bruce, and Loren Lutzenhiser. 1991. 'Social Structures and Economic Conduct: Interpreting Variations in Household Energy Consumption'. *Sociological Forum* 6 (3): 449–70.

Hampton, Sam, and Rob Adams. 2018. 'Behavioural Economics vs Social Practice Theory: Perspectives from inside the United Kingdom Government'. *Energy Research & Social Science* 46 (December): 214–24. https://doi.org/10.1016/j.erss.2018.07.023.

Hansen, Anders Rhiger, Kirsten Gram-Hanssen, and Henrik N. Knudsen. 2018. 'How Building Design and Technologies Influence Heat-Related Habits'. *Building Research & Information* 46 (1): 83–98. https://doi.org/10.1080/09613218.2017.1335477.

Haumont, Nicole. 1968. 'Habitat et modèles culturels'. *Revue française de sociologie* 9 (2): 180–90. https://doi.org/10.2307/3320590.

HCC. 2020. 'Rénover Mieux : Leçons d'Europe'. Réponse à une saisine du gouvernement. Haut Conseil pour le Climat. https://www.hautconseilclimat.fr/wp-content/uploads/2020/11/hcc_rapport_renover_mieux_lecons_deurope.pdf.

Henley, Andrew, and John Peirson. 1998. 'Residential Energy Demand and the Interaction of Price and Temperature: British Experimental Evidence'. *Energy Economics* 20 (2): 157–71. https://doi.org/10.1016/S0140-9883(97)00025-X.

Heydarian, Arsalan, Claire McIlvennie, Laura Arpan, Siavash Yousefi, Marc Syndicus, Marcel Schweiker, Farrokh Jazizadeh, et al. 2020. 'What Drives Our Behaviors in Buildings? A Review

on Occupant Interactions with Building Systems from the Lens of Behavioral Theories'. *Building and Environment* 179 (July): 106928. https://doi.org/10.1016/j.buildenv.2020.106928.

IEA. 2017. 'IEA-EBC Annex 66 - Definition and Simulation of Occupant Behavior in Buildings'. 2017. https://annex66.iea-ebc.org.

Journal Officiel. 2008. *Arrêté Du 8 Août 2008 Portant Approbation de La Méthode de Calcul Th-C-E Ex Prévue Par l'arrêté Du 13 Juin 2008 Relatif à La Performance Énergétique Des Bâtiments Existants de Surface Supérieure à 1 000 Mètres Carrés, Lorsqu'ils Font l'objet de Travaux de Rénovation Importants*. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000019509228.

———. 2021. *Arrêté Du 31 Mars 2021 Relatif Aux Méthodes et Procédures Applicables Au Diagnostic de Performance Énergétique et Aux Logiciels l'établissant*.

Kavgic, M., A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic. 2010. 'A Review of Bottom-up Building Stock Models for Energy Consumption in the Residential Sector'. *Building and Environment* 45 (7): 1683–97. https://doi.org/10.1016/j.buildenv.2010.01.021.

Kuentz-Simonet, Vanessa, Sandrine Lyser, Jacqueline Candau, Philippe Deuffic, Marie Chavent, and Jérôme Saracco. 2013. 'Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement'. *Journal de la société française de statistique* 154 (2): 37–63.

Kuha, Jouni. 2004. 'AIC and BIC: Comparisons of Assumptions and Performance'. *Sociological Methods & Research* 33 (2): 188–229. https://doi.org/10.1177/0049124103262065.

Latour, Bruno. 2007. *Reassembling the Social: An Introduction to Actor-Network-Theory*. OUP Oxford.

Lévy, Jean-Pierre. 1998. 'Habitat et habitants : position et mobilité dans l'espace résidentiel'. In *Trajectoires familiales et espaces de vie en milieu urbain*, 153–80. Transversales. Lyon: Presses universitaires de Lyon. https://doi.org/10.4000/books.pul.9788.

Lévy, Jean-Pierre, and Fateh Belaïd. 2018a. 'Les modèles de consommation énergétique des bâtiments : Limites et perspectives. Chaire Eco-Conception'. Chaire Eco-Conception. ENPC.

———. 2018b. 'The Determinants of Domestic Energy Consumption in France: Energy Modes, Habitat, Households and Life Cycles'. *Renewable and Sustainable Energy Reviews* 81 (January): 2104–14. https://doi.org/10.1016/j.rser.2017.06.022.

Li, Xiwang, and Jin Wen. 2014. 'Review of Building Energy Modeling for Control and Operation'. *Renewable and Sustainable Energy Reviews* 37 (September): 517–37. https://doi.org/10.1016/j.rser.2014.05.056.

Li, Yanfei, Zheng O'Neill, Liang Zhang, Jianli Chen, Piljae Im, and Jason DeGraw. 2021. 'Grey-Box Modeling and Application for Building Energy Simulations - A Critical Review'. *Renewable and Sustainable Energy Reviews* 146 (August): 111174. https://doi.org/10.1016/j.rser.2021.111174.

Liu, Xue, Shan Hu, and Da Yan. 2023. 'A Statistical Quantitative Analysis of the Correlations between Socio-Demographic Characteristics and Household Occupancy Patterns in Residential Buildings in China'. *Energy and Buildings* 284 (April): 112842. https://doi.org/10.1016/j.enbuild.2023.112842.

Loga, Tobias, Britta Stein, and Nikolaus Diefenbach. 2016. 'TABULA Building Typologies in 20 European Countries—Making Energy-Related Features of Residential Building Stocks Comparable'. *Energy and Buildings* 132 (November): 4–12. https://doi.org/10.1016/j.enbuild.2016.06.094.

Lutzenhiser, Loren. 1992. 'A Cultural Model of Household Energy Consumption'. *Energy* 17 (1): 47–60. https://doi.org/10.1016/0360-5442(92)90032-U.

Lutzenhiser, Loren, and Sylvia Bender. 2010. 'The "Average American" Unmasked: Social Structure and Differences in Household Energy Use and Carbon Emissions'. *People-centered initiatives for increasing energy savings*, 191-204.

Madsen, Line Valdorff, and Kirsten Gram-Hanssen. 2017. 'Understanding Comfort and Senses in Social Practice Theory: Insights from a Danish Field Study'. *Energy Research & Social Science* 29 (July): 86–94. https://doi.org/10.1016/j.erss.2017.05.013.

Maresca, Bruno. 2017. 'Mode de vie : de quoi parle-t-on ? Peut-on le transformer ?' *La Pensee ecologique* N° 1 (1): 233–51.

Martin, John Levi. 2018. *Thinking Through Statistics*. Chicago, IL: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/T/bo28394847.html.

Martin, John Levi. 2003. 'What Is Field Theory?' *American Journal of Sociology* 109 (1): 1–49. https://doi.org/10.1086/375201.

Matsumoto, Shigeru. 2016. 'How Do Household Characteristics Affect Appliance Usage? Application of Conditional Demand Analysis to Japanese Household Data'. *Energy Policy* 94 (July): 214–23. https://doi.org/10.1016/j.enpol.2016.03.048.

McKenna, Eoghan, and Murray Thomson. 2016. 'High-Resolution Stochastic Integrated Thermal–Electrical Domestic Demand Model'. *Applied Energy* 165 (March): 445–61. https://doi.org/10.1016/j.apenergy.2015.12.089.

McLoughlin, Fintan, Aidan Duffy, and Michael Conlon. 2015. 'A Clustering Approach to Domestic Electricity Load Profile Characterisation Using Smart Metering Data'. *Applied Energy* 141 (March): 190–99. https://doi.org/10.1016/j.apenergy.2014.12.039.

McMeekin, Andrew, and Dale Southerton. 2012. 'Sustainability Transitions and Final Consumption: Practices and Socio-Technical Systems'. *Technology Analysis & Strategic Management* 24 (4): 345–61. https://doi.org/10.1080/09537325.2012.663960.

MINES ParisTech. 2014. 'COMFIE Simulation Thermique Bâtiments Multizones'. 2014. http://www.ces.mines-paristech.fr/Logiciels/COMFIE/.

Minsky, Marvin. 1965. 'Matter, Mind and Models', March. https://dspace.mit.edu/handle/1721.1/6119.

Monnier, Eric. 1982. 'Les Pratiques Énergétiques Dans l'espace Domestique : Le Cas Des Classes Moyennes En Habitat Collectif Dans La Banlieue Parisienne'. Rapport de recherche 8061383, Paris: Centre de Recherche d'Urbanisme. Ministère de l'Urbanisme et du logement (Plan-Construction).

MTES. 2017. 'Les ménages et la consommation d'énergie', THEMA, 120.

———. 2022a. 'Bilan Énergétique de La France Pour 2020'.

———. 2022b. 'Chiffres Clés Du Climat - France, Europe et Monde'.

Muratori, Matteo, Matthew C. Roberts, Ramteen Sioshansi, Vincenzo Marano, and Giorgio Rizzoni. 2013. 'A Highly Resolved Modeling Technique to Simulate Residential Power Demand'. *Applied Energy* 107 (July): 465–73. https://doi.org/10.1016/j.apenergy.2013.02.057.

Namazkhan, Maliheh, Casper Albers, and Linda Steg. 2020. 'A Decision Tree Method for Explaining Household Gas Consumption: The Role of Building Characteristics, Socio-Demographic Variables, Psychological Factors and Household Behaviour'. *Renewable and Sustainable Energy Reviews* 119 (March): 109542. https://doi.org/10.1016/j.rser.2019.109542.

Nouvel, Romain, Maryam Zirak, Volker Coors, and Ursula Eicker. 2017. 'The Influence of Data Quality on Urban Heating Demand Modeling Using 3D City Models'. *Computers, Environment and Urban Systems* 64 (July): 68–80. https://doi.org/10.1016/j.compenvurbsys.2016.12.005.

Olofsson, T., S. Andersson, and R. Östin. 1998. 'A Method for Predicting the Annual Building Heating Demand Based on Limited Performance Data'. *Energy and Buildings* 28 (1): 101–8. https://doi.org/10.1016/S0378-7788(98)00004-8.

Olu-Ajayi, Razak, Hafiz Alaka, Ismail Sulaimon, Funlade Sunmola, and Saheed Ajayi. 2022. 'Building Energy Consumption Prediction for Residential Buildings Using Deep Learning and Other Machine Learning Techniques'. *Journal of Building Engineering* 45 (January): 103406. https://doi.org/10.1016/j.jobe.2021.103406.

Ortiz, Marco A., and Philomena M. Bluyssen. 2019. 'Developing Home Occupant Archetypes: First Results of Mixed-Methods Study to Understand Occupant Comfort Behaviours and Energy Use in Homes'. *Building and Environment* 163 (October): 106331. https://doi.org/10.1016/j.buildenv.2019.106331.

Pagès, J. 2002. 'Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes'. *Revue de statistique appliquée*, 5–37.

Papineau, Maya, Kareman Yassin, Guy Newsham, and Sarah Brice. 2021. 'Conditional Demand Analysis as a Tool to Evaluate Energy Policy Options on the Path to Grid Decarbonization'. *Renewable and Sustainable Energy Reviews* 149 (October): 111300. https://doi.org/10.1016/j.rser.2021.111300.

Parti, Michael, and Cynthia Parti. 1980. 'The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector'. *The Bell Journal of Economics* 11 (1): 309–21. https://doi.org/10.2307/3003415.

Peuportier, Bruno, and Isabelle Blanc. 1990. 'Simulation Tool with Its Expert Interface for the Thermal Design of Multizone Buildings'. *International Journal of Solar Energy* 8 (2): pages 109-120. https://doi.org/10.1080/01425919008909714.

Pisello, Anna Laura, Veronica Lucia Castaldo, Claudia Fabiani, and Franco Cotana. 2016. 'Investigation on the Effect of Innovative Cool Tiles on Local Indoor Thermal Conditions: Finite Element Modeling and Continuous Monitoring'. *Building and Environment* 97 (February): 55–68. https://doi.org/10.1016/j.buildenv.2015.11.038.

Platten, Jenny von, Mikael Mangold, and Kristina Mjörnell. 2020. 'A Matter of Metrics? How Analysing per Capita Energy Use Changes the Face of Energy Efficient Housing in Sweden and Reveals Injustices in the Energy Transition'. *Energy Research & Social Science* 70 (December): 101807. https://doi.org/10.1016/j.erss.2020.101807.

Qiong Li, Peng Ren, and Qinglin Meng. 2010. 'Prediction Model of Annual Energy Consumption of Residential Buildings'. *2010 International Conference on Advances in Energy Engineering*, June, 223–26. https://doi.org/10.1109/ICAEE.2010.5557576.

Raaij, W. Fred van, and Theo M. M. Verhallen. 1983. 'Patterns of Residential Energy Behavior'. *Journal of Economic Psychology* 4 (1): 85–106. https://doi.org/10.1016/0167-4870(83)90047-8.

Rau, Henrike, Paul Moran, Richard Manton, and Jamie Goggins. 2020. 'Changing Energy Cultures? Household Energy Use before and after a Building Energy Efficiency Retrofit'. *Sustainable Cities and Society* 54 (March): 101983. https://doi.org/10.1016/j.scs.2019.101983.

Raza, Muhammad Qamar, and Abbas Khosravi. 2015. 'A Review on Artificial Intelligence Based Load Demand Forecasting Techniques for Smart Grid and Buildings'. *Renewable and Sustainable Energy Reviews* 50 (October): 1352–72. https://doi.org/10.1016/j.rser.2015.04.065.

Reckwitz, Andreas. 2002. 'Toward a Theory of Social Practices: A Development in Culturalist Theorizing'. *European Journal of Social Theory* 5 (2): 243–63. https://doi.org/10.1177/13684310222225432.

Reynders, Glenn, Thomas Nuytten, and Dirk Saelens. 2013. 'Robustness Of Reduced-Order Models For Prediction And Simulation Of The Thermal Behavior Of Dwellings'. In *Proceedings of*

BS2013: 13th conference of international building performance simulation association, Chambéry, France.

Richardson, Ian, Murray Thomson, and David Infield. 2008. 'A High-Resolution Domestic Building Occupancy Model for Energy Demand Simulations'. *Energy and Buildings* 40 (8): 1560–66. https://doi.org/10.1016/j.enbuild.2008.02.006.

Risch, Anna, and Claire Salmon. 2017. 'What Matters in Residential Energy Consumption: Evidence from France'. *International Journal of Global Energy Issues* 40 (1–2): 79–116. https://doi.org/10.1504/IJGEI.2017.080767.

Runge, Jason, and Radu Zmeureanu. 2019. 'Forecasting Energy Use in Buildings Using Artificial Neural Networks: A Review'. *Energies* 12 (17): 3254. https://doi.org/10.3390/en12173254.

Salvó, G., and M. N. Piacquadio. 2017. 'Multifractal Analysis of Electricity Demand as a Tool for Spatial Forecasting'. *Energy for Sustainable Development* 38 (June): 67–76. https://doi.org/10.1016/j.esd.2017.02.005.

Saporta, Gilbert. 2006. *Probabilités, analyse des données et statistique*. Editions TECHNIP.

Sekhar Roy, Sanjiban, Reetika Roy, and Valentina E. Balas. 2018. 'Estimating Heating Load in Buildings Using Multivariate Adaptive Regression Splines, Extreme Learning Machine, a Hybrid Model of MARS and ELM'. *Renewable and Sustainable Energy Reviews* 82 (February): 4256–68. https://doi.org/10.1016/j.rser.2017.05.249.

Selosse, Margot, Julien Jacques, and Christophe Biernacki. 2020. 'Model-Based Co-Clustering for Mixed Type Data'. *Computational Statistics & Data Analysis* 144 (April): 106866. https://doi.org/10.1016/j.csda.2019.106866.

Sergent, Michelle, Didier Mathieu, Roger Phan-Tan-Luu, and Giuliana Drava. 1995. 'Correct and Incorrect Use of Multilinear Regression'. *Chemometrics and Intelligent Laboratory Systems* 27 (2): 153–62. https://doi.org/10.1016/0169-7439(95)80020-A.

Serrano-Jiménez, Antonio, Paula Femenías, Liane Thuvander, and Ángela Barrios-Padura. 2021. 'A Multi-Criteria Decision Support Method towards Selecting Feasible and Sustainable Housing Renovation Strategies'. *Journal of Cleaner Production* 278 (January): 123588. https://doi.org/10.1016/j.jclepro.2020.123588.

Shove, Elizabeth. 2003. 'Converging Conventions of Comfort, Cleanliness and Convenience'. *Journal of Consumer Policy* 26 (4): 395–418. https://doi.org/10.1023/A:1026362829781.

———. 2017. 'Energy and Social Practice: From Abstractions to Dynamic Processes'. In *Complex Systems and Social Practices in Energy Transitions*, edited by Nicola Labanca, 207–20. Green Energy and Technology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33753-1_9.

Shove, Elizabeth, Mika Pantzar, and Matt Watson. 2012. *The Dynamics of Social Practice: Everyday Life and How It Changes*. SAGE.

Shove, Elizabeth, Gordon Walker, and Sam Brown. 2014. 'Transnational Transitions: The Diffusion and Integration of Mechanical Cooling'. *Urban Studies* 51 (7): 1506–19. https://doi.org/10.1177/0042098013500084.

Spurling, Nicola Jane, Andrew McMeekin, Dale Southerton, Elizabeth Anne Shove, and Daniel Welch. 2013. *Interventions in Practice: Reframing Policy Approaches to Consumer Behaviour*. Manchester: Sustainable Practices Research Group. https://eprints.lancs.ac.uk/id/eprint/85608/.

Stephenson, Janet, Barry Barton, Gerry Carrington, Daniel Gnoth, Rob Lawson, and Paul Thorsnes. 2010. 'Energy Cultures: A Framework for Understanding Energy Behaviours'. *Energy Policy*, The socio-economic transition towards a hydrogen economy - findings from European research, with regular papers, 38 (10): 6120–29. https://doi.org/10.1016/j.enpol.2010.05.069.

Stock, Mathis. 2003. ‘Pratiques des lieux, modes d’habiter, régimes d’habiter : Pour une analyse trialogique des dimensions spatiales des sociétés humaines’. *Travaux de l’Institut Géographique de Reims* 29 (115): 213–29. https://doi.org/10.3406/tigr.2003.1473.

Subrémon, Hélène. 2009. ‘Habiter Avec l’énergie. Pour Une Anthropologie Sensible de La Consommation d’énergie’. These de doctorat, Paris 10. https://www.theses.fr/2009PA100039.

Sütterlin, Bernadette, Thomas A. Brunner, and Michael Siegrist. 2011. ‘Who Puts the Most Energy into Energy Conservation? A Segmentation of Energy Consumers Based on Energy-Related Behavioral Characteristics’. *Energy Policy*, Clean Cooking Fuels and Technologies in Developing Economies, 39 (12): 8137–52. https://doi.org/10.1016/j.enpol.2011.10.008.

Swan, Lukas G., and V. Ismet Ugursal. 2009. ‘Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques’. *Renewable and Sustainable Energy Reviews* 13 (8): 1819–35. https://doi.org/10.1016/j.rser.2008.09.033.

Taleghani, Mohammad, Martin Tenpierik, Stanley Kurvers, and Andy van den Dobbelsteen. 2013. ‘A Review into Thermal Comfort in Buildings’. *Renewable and Sustainable Energy Reviews* 26 (October): 201–15. https://doi.org/10.1016/j.rser.2013.05.050.

Tanimoto, Jun, Aya Hagishima, and Hiroki Sagara. 2008. ‘A Methodology for Peak Energy Requirement Considering Actual Variation of Occupants’ Behavior Schedules’. *Building and Environment*, Part Special: Building Performance Simulation, 43 (4): 610–19. https://doi.org/10.1016/j.buildenv.2006.06.034.

Tsanas, Athanasios, and Angeliki Xifara. 2012. ‘Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools’. *Energy and Buildings* 49 (June): 560–67. https://doi.org/10.1016/j.enbuild.2012.03.003.

Tso, Geoffrey K. F., and Kelvin K. W. Yau. 2007. ‘Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks’. *Energy* 32 (9): 1761–68. https://doi.org/10.1016/j.energy.2006.11.010.

Van Raaij, W. Fred, and Theo M. M. Verhallen. 1983. ‘A Behavioral Model of Residential Energy Use’. *Journal of Economic Psychology* 3 (1): 39–63. https://doi.org/10.1016/0167-4870(83)90057-0.

Varenne, Franck. 2008. ‘Epistémologie des modèles et des simulations’. In *Les modèles, possibilités et limites* (pp. 13-46). Editions Matériologiques.

Varenne, Franck, and Marc Silberstein. 2013. *Modéliser & simuler. Epistémologies et pratiques de la modélisation et de la simulation, tome 1*. Vol. 1. Editions Matériologiques. https://hal.inria.fr/hal-00826655.

Verbeke, Stijn, and Amaryllis Audenaert. 2018. ‘Thermal Inertia in Buildings: A Review of Impacts across Climate and Building Use’. *Renewable and Sustainable Energy Reviews* 82 (February): 2300–2318. https://doi.org/10.1016/j.rser.2017.08.083.

Vlek, B. Gatersleben, Ch. 1998. ‘Household Consumption, Quality of Life, and Environmental Impacts: A Psychological Perspective and Empirical Study’. In *Green Households*. Routledge.

Vorger, Éric. 2014. ‘Étude de l’influence Du Comportement Des Habitants Sur La Performance Énergétique Du Bâtiment’. These de doctorat, Paris, ENMP. http://www.theses.fr/2014ENMP0066.

Wang, Zequn, and Yuxiang Chen. 2019. ‘Data-Driven Modeling of Building Thermal Dynamics: Methodology and State of the Art’. *Energy and Buildings* 203 (November): 109405. https://doi.org/10.1016/j.enbuild.2019.109405.

Warriner, G. Keith, Gordon H. G. McDougall, and John D. Claxton. 1984. ‘Any Data or None at All?: Living with Inaccuracies in Self-Reports of Residential Energy Consumption’. *Environment and Behavior* 16 (4): 503–26. https://doi.org/10.1177/0013916584164005.

Widén, Joakim, Andreas Molin, and Kajsa Ellegård. 2012. 'Models of Domestic Occupancy, Activities and Energy Use Based on Time-Use Data: Deterministic and Stochastic Approaches with Application to Various Building-Related Simulations'. *Journal of Building Performance Simulation* 5 (1): 27–44. https://doi.org/10.1080/19401493.2010.532569.

Wilhite, Harold, Hidetoshi Nakagami, Takashi Masuda, Yukiko Yamaga, and Hiroshi Haneda. 1996. 'A Cross-Cultural Analysis of Household Energy Use Behaviour in Japan and Norway'. *Energy Policy* 24 (9): 795–803. https://doi.org/10.1016/0301-4215(96)00061-4.

Wise, Alyssa Friend, and David Williamson Shaffer. 2015. 'Why Theory Matters More than Ever in the Age of Big Data'. *Journal of Learning Analytics* 2 (2): 5–13. https://doi.org/10.18608/jla.2015.22.2.

Wu, Shimei, Xinye Zheng, Jin Guo, Chuan-Zhong Li, and Chu Wei. 2020. 'Quantifying Energy Consumption in Household Surveys: An Alternative Device-Based Accounting Approach'. *Field Methods* 32 (2): 213–32. https://doi.org/10.1177/1525822X20905790.

Zhao, Dong, Andrew P. McCoy, Jing Du, Philip Agee, and Yujie Lu. 2017. 'Interaction Effects of Building Technology and Resident Behavior on Energy Consumption in Residential Buildings'. *Energy and Buildings* 134 (January): 223–33. https://doi.org/10.1016/j.enbuild.2016.10.049.

Annexes

Annexe 1 : Tableaux de données résumant les résultats de classification des données de comportement (ENERGIHAB)

Typologie 1 - GOWER : Classification ascendante hiérarchique à partir des distances de Gower entre les lignes de la base de données (voir Chapitre 2).

	GOW_1	GOW_2	GOW_3	GOW_4	GOW_5	GOW_6	GOW_7	Total
Age de la PR								
Moins de 30 ans	**6%**	4%	2%	1%	2%	1%	**5%**	3%
30-39 ans	15%	16%	16%	2%	17%	11%	23%	13%
40-49 ans	19%	28%	31%	5%	23%	21%	34%	22%
50-59 ans	21%	26%	38%	8%	21%	18%	26%	23%
60-69 ans	20%	15%	10%	24%	19%	21%	10%	17%
70+ ans	19%	11%	3%	60%	17%	28%	3%	22%
Type de logement								
Locataire (secteur privé)	32%	31%	14%	30%	29%	14%	25%	25%
Locataire (secteur public)	14%	18%	8%	12%	14%	13%	19%	13%
Logé gratuitement	2%	6%	3%	1%	3%	3%	5%	3%
Propriétaires (dont accédants)	52%	46%	75%	57%	54%	70%	51%	58%
Composition du ménage								
Couple avec enfants	15%	32%	64%	8%	29%	44%	62%	33%
Couple sans enfants	26%	19%	22%	31%	26%	36%	21%	26%
Famille monoparentale	9%	11%	6%	3%	7%	6%	6%	7%
Personne seule	46%	36%	7%	57%	37%	9%	10%	32%
Autre	3%	2%	2%	0%	1%	4%	2%	2%
Nombre de personnes								
1 pers.	46%	36%	7%	57%	37%	9%	10%	32%
2 pers.	36%	26%	26%	34%	31%	41%	24%	31%
3 pers.	7%	18%	20%	3%	10%	16%	30%	13%
4 pers.	8%	14%	31%	4%	15%	19%	27%	16%
5 pers.	2%	6%	12%	1%	5%	9%	6%	6%
6+ pers.	1%	1%	4%	1%	1%	5%	3%	2%
Catégorie socio-professionnelle								
Employés et ouvriers	27%	33%	31%	50%	33%	34%	50%	36%
Cadres	26%	26%	25%	15%	18%	30%	18%	23%
Agriculteurs et artisans	6%	6%	4%	7%	4%	2%	5%	5%
Professions intermédiaires	40%	33%	40%	23%	40%	31%	27%	34%
Sans profession	1%	2%	0	5%	4%	2%	0	2%
Revenus								
Bas revenus	50%	41%	12%	68%	55%	23%	22%	40%
Revenus moyens	30%	30%	36%	23%	21%	34%	43%	30%
Revenus élevés	21%	30%	52%	9%	24%	43%	35%	30%
Statut d'activité de la PR								
Actif	61%	72%	86%	6%	53%	32%	79%	54%
Autre	2%	4%	3%	8%	8%	13%	3%	6%
Retraité	32%	19%	9%	82%	31%	47%	13%	35%
Etudiant	0	2%	0%	0	0	0	2%	1%
Sans emploi	5%	3%	2%	5%	8%	7%	3%	4%
Zone urbaine								
Zone périurbaine	9%	12%	7%	10%	13%	7%	5%	9%
Zone rurale	37%	27%	52%	34%	44%	30%	49%	38%
Zone urbaine	54%	61%	41%	57%	43%	63%	46%	52%
Surface habitable du logement								
<30m²	11%	14%	34%	12%	14%	31%	19%	19%

30-50m²	25%	16%	4%	23%	19%	4%	10%	15%
50-75m²	29%	31%	15%	29%	27%	18%	24%	25%
75-100m²	23%	27%	30%	24%	28%	33%	37%	28%
100-150m²	8%	7%	0	7%	5%	1%	2%	5%
>150m²	4%	5%	17%	4%	6%	14%	10%	8%
Surface par personne								
Moyenne	46,7	43,8	42,2	54,1	45,5	44,3	36,1	45,8
Ecart-type	26,5	30,3	31,1	29,6	28,2	27,9	23,6	29,2
Logement collectif ou individuel								
Collectif	67%	70%	36%	64%	60%	48%	54%	57%
Individuel	33%	30%	64%	36%	40%	52%	46%	43%
Chauffage central								
Oui	51%	65%	57%	62%	55%	66%	60%	59%
Non	49%	35%	43%	38%	45%	34%	40%	41%
Type d'énergie utilisée pour le chauffage principal								
Electricité	35%	35%	32%	32%	37%	39%	24%	34%
Fioul	7%	8%	5%	5%	7%	7%	6%	6%
Gaz	41%	47%	49%	48%	42%	42%	56%	46%
Autre	17%	11%	14%	15%	15%	12%	14%	14%
Attitude vis-à-vis de la consommation d'énergie actuelle								
"Gros consommateur"	7%	8%	9%	9%	7%	10%	**13%**	9%
"Consommateur moyen"	58%	60%	61%	62%	62%	57%	**66%**	61%
"Petit consommateur"	**34%**	32%	30%	29%	31%	32%	21%	31%
Température de chauffe idéale								
Moyenne	20,4	20,6	20,3	20,4	20,5	20,5	20,3	20,4
Ecart type	1,8	2,5	1,6	2,0	1,6	1,9	1,9	1,9
Logement considéré comme confortable								
Oui	92%	94%	96%	94%	93%	93%	95%	94%
Non	**8%**	6%	4%	6%	7%	7%	5%	6%
Durée d'occupation du logement								
Mean duration (years)	16,3	15,4	**13,3**	**27,3**	15,4	**19,9**	**13,0**	17,6
SD of duration (years)	12,4	12,8	9,7	18,1	13,1	14,5	9,7	14,3
Motivation du dernier déménagement								
Distance	9%	7%	7%	8%	5%	**11%**	8%	8%
Cadre de vie	**30%**	22%	22%	**35%**	23%	22%	29%	26%
Besoins liés au ménage	31%	28%	28%	21%	31%	**32%**	29%	28%
Dernier logement plus disponible	4%	8%	9%	7%	7%	**13%**	6%	8%
Autre	6%	**10%**	7%	5%	8%	7%	**13%**	7%
Accès à la propriété	8%	9%	11%	12%	11%	11%	10%	10%
Prix	4%	3%	2%	3%	2%	1%	0%	3%
Travail	9%	**14%**	10%	7%	9%	9%	5%	10%
Département								
75 - Paris	15%	**22%**	10%	**21%**	13%	20%	8%	16%
77 - Seine et Marne	**15%**	8%	9%	13%	13%	10%	**14%**	11%
78 - Yvelines	11%	12%	**17%**	13%	14%	10%	11%	13%
91 - Essone	11%	9%	**21%**	11%	13%	6%	11%	12%
92 - Hauts de Seine	11%	**15%**	9%	12%	12%	16%	6%	12%
93 - Seine-Saint-Denis	14%	13%	9%	9%	9%	14%	**16%**	11%
94 - Val-de-Marne	14%	12%	13%	15%	8%	12%	**16%**	13%
95 - Val d'Oise	9%	10%	13%	7%	**16%**	11%	**17%**	11%

Typologie 2 – AFDM

	AFDM_1	AFDM_2	AFDM_3	AFDM_4	AFDM_5	AFDM_6	AFDM_7	Total
Age de la PR								
Moins de 30 ans	1%	**5%**	**6%**	3%	1%	3%	1%	3%
30-39 ans	1%	**22%**	12%	**21%**	4%	**26%**	12%	13%
40-49 ans	5%	20%	23%	24%	13%	**31%**	**40%**	22%
50-59 ans	12%	**28%**	23%	26%	13%	26%	**33%**	23%
60-69 ans	**22%**	16%	16%	17%	**27%**	11%	11%	17%
70+ ans	**59%**	9%	19%	9%	**42%**	3%	2%	22%
Type de logement								
Locataire (secteur privé)	**30%**	**46%**	25%	**30%**	14%	25%	12%	25%
Locataire (secteur public)	17%	14%	**18%**	**18%**	5%	10%	10%	13%
Logé gratuitement	2%	4%	4%	4%	1%	3%	4%	3%
Propriétaires (dont accédants)	50%	36%	53%	48%	**81%**	**62%**	**75%**	58%
Composition du ménage								
Couple avec enfants	5%	8%	31%	36%	30%	**44%**	**75%**	33%
Couple sans enfants	24%	13%	25%	25%	**48%**	29%	16%	26%
Famille monoparentale	4%	5%	**13%**	**11%**	4%	8%	5%	7%
Personne seule	**65%**	**72%**	29%	27%	17%	17%	3%	32%
Autre	2%	2%	2%	2%	2%	3%	2%	2%
Nombre de personnes								
1 pers.	**65%**	**72%**	29%	27%	17%	17%	3%	32%
2 pers.	29%	19%	34%	32%	**52%**	35%	17%	31%
3 pers.	3%	5%	**18%**	**17%**	10%	**19%**	**19%**	13%
4 pers.	2%	2%	13%	17%	13%	**23%**	**36%**	16%
5 pers.	2%	1%	2%	6%	6%	6%	**19%**	6%
6+ pers.	1%	1%	3%	2%	2%	0%	**7%**	2%
Catégorie socio-professionnelle								
Employés et ouvriers	**53%**	32%	41%	39%	28%	21%	29%	36%
Cadres	13%	24%	26%	18%	26%	27%	29%	23%
Agriculteurs et artisans	7%	**8%**	4%	3%	**9%**	4%	3%	5%
Professions intermédiaires	21%	36%	29%	38%	34%	**46%**	39%	34%
Sans profession	**6%**	1%	1%	2%	**4%**	1%	0%	2%
Revenus								
Bas revenus	**77%**	**60%**	45%	43%	33%	19%	10%	40%
Revenus moyens	18%	28%	35%	**39%**	35%	31%	26%	30%
Revenus élevés	5%	12%	21%	19%	33%	**50%**	**65%**	30%
Statut d'activité de la PR								
Actif	10%	**73%**	58%	**69%**	15%	**87%**	**81%**	54%
Autre	6%	4%	7%	4%	**12%**	2%	5%	6%
Retraité	**77%**	18%	30%	23%	**68%**	7%	10%	35%
Etudiant	0%	2%	1%	0%	0%	0%	1%	1%
Sans emploi	**7%**	3%	4%	4%	6%	3%	3%	4%
Zone urbaine								
Zone périurbaine	**14%**	**15%**	10%	11%	8%	4%	5%	9%
Zone rurale	31%	26%	34%	**45%**	38%	42%	**50%**	38%
Zone urbaine	55%	**58%**	56%	44%	54%	54%	46%	52%
Surface habitable du logement								
<30m²	9%	2%	15%	13%	**30%**	21%	**41%**	19%
30-50m²	26%	**41%**	14%	17%	5%	11%	0%	15%
50-75m²	**31%**	29%	**34%**	29%	17%	26%	8%	25%
75-100m²	24%	8%	28%	32%	**33%**	33%	30%	28%
100-150m²	**9%**	**20%**	5%	2%	0%	1%	0%	5%
>150m²	1%	0%	4%	7%	**16%**	7%	**21%**	8%
Surface par personne								
Moyenne	**53,5**	41,2	42,8	43,4	**55,7**	41,1	40,3	45,8
Ecart-type	29,0	23,1	27,7	29,3	34,5	25,9	28,8	29,2
Logement collectif ou individuel								
Collectif	**69%**	**82%**	**69%**	61%	39%	57%	32%	57%

Individuel	31%	18%	31%	39%	**61%**	43%	**68%**	43%
Chauffage central								
Oui	60%	51%	**65%**	60%	**65%**	55%	58%	59%
Non	40%	49%	35%	40%	35%	45%	42%	41%
Type d'énergie utilisée pour le chauffage principal								
Electricité	34%	**41%**	35%	32%	32%	32%	33%	34%
Fioul	6%	**9%**	6%	4%	7%	8%	6%	6%
Gaz	47%	37%	43%	**50%**	45%	48%	46%	46%
Autre	13%	14%	16%	15%	16%	11%	15%	14%
Attitude vis-à-vis de la consommation d'énergie actuelle								
"Gros consommateur"	7%	7%	10%	11%	6%	7%	10%	9%
"Consommateur moyen"	63%	55%	61%	60%	56%	63%	63%	61%
"Petit consommateur"	30%	**38%**	28%	29%	**38%**	30%	27%	31%
Température de chauffe idéale								
Moyenne	20,3	20,5	20,4	20,5	20,6	20,5	20,3	20,4
Ecart type	1,8	1,7	2,4	2,0	2,2	1,7	1,5	1,9
Logement considéré comme confortable								
Oui	92%	93%	93%	**96%**	94%	94%	95%	94%
Non	**8%**	7%	7%	4%	6%	6%	5%	6%
Durée d'occupation du logement								
Durée moyenne (années)	**26,8**	14,0	15,8	14,3	**24,8**	11,9	13,3	17,6
Écart type de la durée (années)	18,3	10,4	12,1	11,9	16,3	9,2	8,3	14,3
Motivation du dernier déménagement								
Distance	9%	7%	7%	8%	5%	**11%**	8%	8%
Cadre de vie	**33%**	**36%**	22%	23%	26%	20%	23%	26%
Besoins liés au ménage	27%	22%	32%	**32%**	28%	**31%**	25%	28%
Dernier logement plus disponible	8%	6%	9%	5%	6%	**10%**	**10%**	8%
Autre	5%	6%	7%	**11%**	7%	6%	**10%**	7%
Accès à la propriété	8%	10%	9%	9%	**13%**	8%	**15%**	10%
Prix	2%	4%	2%	3%	**5%**	2%	1%	3%
Travail	7%	10%	11%	10%	11%	**12%**	8%	10%
Département								
75 - Paris	19%	**23%**	17%	11%	18%	19%	10%	16%
77 - Seine et Marne	**14%**	13%	11%	13%	11%	8%	9%	11%
78 - Yvelines	11%	12%	10%	14%	14%	13%	**17%**	13%
91 - Essonne	11%	8%	11%	12%	**15%**	14%	14%	12%
92 - Hauts de Seine	11%	14%	13%	11%	**15%**	10%	13%	12%
93 - Seine-Saint-Denis	12%	14%	13%	11%	6%	13%	10%	11%
94 - Val-de-Marne	13%	8%	13%	12%	**15%**	13%	13%	13%
95 - Val d'Oise	9%	9%	12%	**17%**	6%	11%	**15%**	11%

Tableau 26 : Tableau de synthèse de l'AFDM sur les données de comportements. La partie gauche recense les contributions des modalités à la constructions des variables synthétiques (en %). La partie droite recense la qualité de la projection des mêmes modalités sur chacun des axes principaux. Une contribution supérieure à 5% est statistiquement significative. Un cos² proche de 1 signifie que la modalité est bien représentée sur l'axe factoriel (et donc que la modalité peu aider à interpréter le sens de celui-ci). Les cases sont colorées pour mettre en avant les modalités contributrices et bien représentées. Seuls les 10 premiers axes sont représentés.

		Contribution des variables (%)										Qualité de la projection (cos²)									
Variable	Modalité	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Variables qualitatives																					
F_EQ2	IND.FREEZ_Y	2,3	1,7	1,4	0,8	0,4	0,0	0,7	8,9	0,1	6,4	0,4	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,1
	IND.FREEZ_N	1,4	1,0	0,8	0,4	0,2	0,0	0,4	5,2	0,1	3,7	0,4	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,1
HY_US1	<1SHOWER	0,0	0,4	0,0	0,2	0,1	0,2	6,8	1,7	11,2	3,9	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,2	0,1
	=1SHOWER	0,0	0,0	1,3	0,0	0,9	0,7	0,7	0,1	4,6	0,0	0,0	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,2	0,0
	>1SHOWER	0,1	1,0	7,1	0,6	4,0	2,1	1,8	0,6	0,6	5,1	0,0	0,0	0,2	0,0	0,1	0,0	0,0	0,0	0,0	0,1
HY_US2	BATH_N	0,0	0,0	0,0	0,0	0,2	0,1	0,1	0,2	0,8	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,2	0,0
	BATH_Y	0,1	0,4	0,1	0,0	3,7	1,3	1,3	4,5	18,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,2	0,0
HY_REG1	WATERSAV_N	0,0	0,3	0,1	0,0	0,3	1,6	0,0	0,2	0,5	0,3	0,0	0,1	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0
	WATERSAV_Y	0,0	2,4	0,9	0,3	2,3	11,3	0,2	1,2	3,3	2,0	0,0	0,1	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0
HY_REG2	SELECSORT_N	0,4	0,0	3,0	1,7	0,5	1,1	6,4	1,5	14,1	3,4	0,1	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,2	0,0
	SELECSORT_Y	0,0	0,0	0,3	0,2	0,1	0,1	0,7	0,2	1,4	0,3	0,1	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,2	0,0
HY_REG3	GREENPROD_N	0,1	0,6	2,2	3,7	1,2	0,7	5,8	2,2	0,0	1,1	0,0	0,0	0,1	0,1	0,0	0,0	0,1	0,0	0,0	0,0
	GREENPROD_Y	0,1	0,4	1,4	2,4	0,7	0,4	3,7	1,4	0,0	0,7	0,0	0,0	0,1	0,1	0,0	0,0	0,1	0,0	0,0	0,0
LI_REG1	LEDBC_N	1,7	0,5	2,0	6,7	0,7	0,3	0,5	0,8	1,0	0,1	0,3	0,0	0,1	0,2	0,0	0,0	0,0	0,0	0,0	0,0
	LEDBC_SMT	0,1	0,8	1,6	4,1	0,3	0,1	3,1	0,0	0,3	0,4	0,0	0,0	0,1	0,2	0,0	0,0	0,1	0,0	0,0	0,0
	LEDBC_ALW	0,7	2,6	0,0	21,0	0,0	0,6	1,6	0,7	0,1	1,1	0,1	0,1	0,0	0,6	0,0	0,0	0,0	0,0	0,0	0,0
LI_REG2	HALOG_N	0,2	3,6	0,2	25,1	0,0	0,7	2,5	0,8	0,7	0,3	0,0	0,1	0,0	0,6	0,0	0,0	0,0	0,0	0,0	0,0
	HALOG_Y	0,1	1,1	0,1	7,7	0,0	0,2	0,8	0,3	0,2	0,1	0,0	0,1	0,0	0,6	0,0	0,0	0,0	0,0	0,0	0,0
LI_REG3	REGLUX_N	0,3	0,0	0,0	2,7	0,3	8,9	2,5	0,0	4,9	1,6	0,1	0,0	0,0	0,1	0,0	0,2	0,0	0,0	0,1	0,0
	REGLUX_Y	0,0	0,0	0,0	0,3	0,0	0,9	0,3	0,0	0,5	0,2	0,1	0,0	0,0	0,1	0,0	0,2	0,0	0,0	0,1	0,0
TC_EQ1	AUXHEAT_N	0,0	0,2	0,6	0,1	0,8	2,4	0,1	2,4	0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,1	0,0	0,0
	AUXHEAT_Y	0,1	0,5	1,4	0,1	2,0	5,9	0,2	6,0	0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,1	0,0	0,0
TC_REG1	LIGHTCLOTH	0,2	0,5	8,9	0,7	0,2	0,8	0,6	0,7	0,6	0,0	0,0	0,0	0,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	INTERMCLOTH	0,0	0,0	0,2	0,1	6,0	3,0	5,6	4,1	0,3	2,7	0,0	0,0	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0
	WARMCLOTH	0,1	0,3	5,5	0,1	1,7	0,3	1,1	0,6	0,0	1,2	0,0	0,0	0,3	0,0	0,1	0,0	0,0	0,0	0,0	0,0
TC_REG2	REGULT_N	0,7	0,5	5,2	0,0	0,2	0,6	0,0	0,0	0,7	2,8	0,2	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,1
	REGULT_Y	0,5	0,4	3,9	0,0	0,1	0,4	0,0	0,0	0,5	2,1	0,2	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,1
TC_REG4	HEATOFF_N	0,1	0,6	3,6	0,5	0,0	1,6	2,4	0,1	0,0	1,2	0,0	0,0	0,2	0,0	0,0	0,1	0,1	0,0	0,0	0,0
	HEATOFF_Y	0,1	0,7	4,0	0,6	0,0	1,8	2,7	0,2	0,0	1,3	0,0	0,0	0,2	0,0	0,0	0,1	0,1	0,0	0,0	0,0
WL_REG1	SCREENREG_LOW	5,3	2,5	0,5	0,8	5,6	0,2	0,5	3,3	0,5	0,8	0,7	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
	SCREENREG_MID	0,4	0,4	0,3	0,0	1,9	2,1	0,7	0,0	0,2	7,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1
	SCREENREG_HIGH	2,5	0,7	1,7	0,8	0,8	1,1	0,0	2,9	1,3	3,5	0,5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
OCC1	OCCWE_LOW	0,1	6,9	0,1	0,8	0,0	0,8	1,6	0,6	0,0	2,9	0,0	0,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	OCCWE_MID	0,3	0,3	0,1	0,2	0,6	0,0	1,0	0,0	1,8	5,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1
	OCCWE_HIGH	0,1	2,2	0,0	1,2	0,8	0,5	0,0	0,2	1,6	1,2	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
OCC2	VAC_N	2,6	1,7	3,0	0,3	0,0	0,7	7,1	0,2	0,0	0,0	0,4	0,1	0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,0
	VAC_ST	0,3	0,2	0,3	0,1	0,0	0,0	7,0	0,5	0,0	0,1	0,2	0,0	0,0	0,0	0,0	0,0	0,3	0,0	0,0	0,0
	VAC_OFT	1,0	0,9	1,8	2,0	0,1	1,6	5,7	1,1	0,1	0,3	0,2	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0
OCC3	AWAY_EVD	0,9	8,4	0,6	0,0	0,5	0,4	0,8	2,1	0,2	0,0	0,2	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	AWAY_ST	0,7	0,5	3,2	0,5	0,0	0,1	2,6	8,3	0,3	0,5	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,1	0,0	0,0
	HOMEBASED	1,7	10,3	0,0	0,3	0,6	0,5	0,0	0,0	0,1	0,0	0,4	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
TC_US3	VENTILBLOCK_Y	0,0	0,3	1,1	0,3	0,1	10,4	0,0	0,9	2,0	9,9	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,1
	VENTILBLOCK_N	0,0	0,0	0,0	0,0	0,0	0,4	0,0	0,0	0,1	0,3	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,1
TC_US4	REGULT_WINDOWS_N	0,0	0,0	0,0	0,0	0,8	0,0	0,1	0,1	0,3	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
	REGULT_WINDOWS_Y	0,0	0,2	0,0	0,1	5,7	0,1	0,4	0,4	2,1	0,1	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
Variables quantitatives																					
F_EQ1		9,2	3,1	0,6	0,6	0,2	0,1	0,0	7,4	0,3	3,4	0,4	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0
F_US1		0,5	19,8	0,1	2,7	3,0	0,1	1,2	6,5	0,8	0,5	0,0	0,4	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0
F_US2		1,2	0,9	0,1	0,0	0,8	7,1	3,6	7,8	6,3	0,9	0,1	0,0	0,0	0,0	0,0	0,1	0,0	0,1	0,1	0,0
F_US3		1,9	1,7	0,7	0,5	4,7	3,3	1,1	0,4	1,7	3,4	0,1	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
F_REG1		5,2	0,2	3,0	0,7	0,0	0,3	1,3	3,4	1,6	2,1	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
HY_EQ1		12,8	4,4	0,1	0,3	2,3	0,4	0,1	0,1	0,1	1,0	0,6	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
HY_US3		10,8	3,0	0,5	0,8	3,5	0,0	0,0	0,2	0,1	1,3	0,5	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
HY_US4		7,4	1,2	0,2	0,0	3,1	0,1	0,1	0,6	0,0	0,0	0,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
HY_US5		6,5	1,4	0,8	0,8	4,5	0,0	0,9	0,1	1,1	1,0	0,3	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
LI_EQ1		0,5	1,6	2,2	1,7	0,4	14,3	1,0	0,1	0,7	3,6	0,0	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0
TC_US1		0,0	0,0	12,7	0,6	0,2	5,3	0,0	2,3	2,5	0,6	0,0	0,0	0,2	0,0	0,0	0,1	0,0	0,0	0,0	0,0
TC_US2		0,0	2,6	2,2	2,2	1,6	0,0	9,2	0,1	0,0	6,3	0,0	0,1	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1
TC_REG3		0,0	0,1	4,7	0,8	0,4	0,2	0,5	2,1	8,0	0,2	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,0
WL_EQ1		13,4	1,5	1,0	0,5	2,3	0,2	0,0	0,6	0,0	0,1	0,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
WL_US1		0,5	0,0	1,8	0,2	20,9	1,2	1,2	0,9	0,2	1,2	0,0	0,0	0,0	0,0	0,3	0,0	0,0	0,0	0,0	0,0
WL_US2		4,8	2,1	1,1	0,2	7,7	0,0	0,0	2,1	0,9	0,6	0,2	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0

Typologie 3 – COV

Tableau 27 : Position des modalités des variables qualitatives et corrélation des variables quantitatives dans l'espace synthétique. Source : Auteur, calculs réalisés sur la base ENERGIHAB.

Variable synthétique	Variable de comportement	Corrélation	Coordonnées de la modalité
VS1 [-2,67 ; 3,26]	F_EQ1	-0,86	
	F_US3	-0,49	
	F_EQ2=INDFREEZ_N		0,48

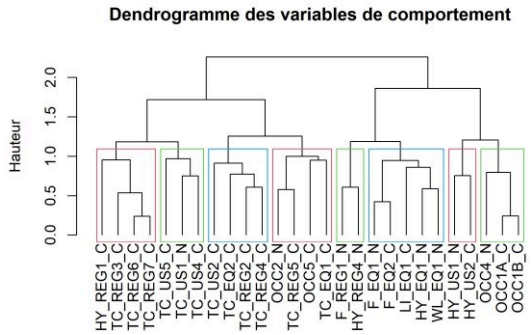
	F_EQ2=INDFREEZ_Y		-0,82
VS2 [-1,33 ; 2,30]	F_US1	-0,80	
	OCC3=OCCWE_HIGH		-0,37
	OCC3=OCCWE_LOW		0,78
	OCC3=OCCWE_MID		-0,07
	OCC5=AWAY_EVD		0,63
	OCC5=AWAY_ST		0,25
	OCC5=HOMEBASED		-0,68
VS3 [-3,36 ; 2,42]	F_US2	-0,28	
	F_REG1	-0,50	
	HY_EQ1	-0,84	
	HY_US3	-0,79	
	HY_US4	-0,65	
	HY_US5	-0,63	
	LI_REG3=REGLUX_N		-0,29
	LI_REG3=REGLUX_Y		0,03
	OCC4=VAC_N		0,41
	OCC4=VAC_OFT		-0,41
	OCC4=VAC_ST		-0,09
VS4 [-1,42 ; 6,07]	LI_EQ1	0,51	
	HY_REG1=WATERSAV_N		-0,20
	HY_REG1=WATERSAV_Y		1,37
	TC_EQ1=AUXHEAT_N		-0,32
	TC_EQ1=AUXHEAT_Y		0,78
	TC_USE3=VENTILBLOCK_N		-0,09
	TC_USE3=VENTILBLOCK_Y		2,72
VS5 [-2,12 ; 5,77]	TC_US1	0,66	
	HY_US1=<1SHOWER		-0,19
	HY_US1==1SHOWER		-0,21
	HY_US1=>1SHOWER		1,46
	TC_REG1=INTERMCLOTH		0,19
	TC_REG1=LIGHTCLOTH		0,90
	TC_REG1=WARMCLOTH		-0,52
VS6 [-1,46 ; 3,03]	TC_US2	-0,49	
	TC_REG3	0,38	
	TC_REG2=REGULT_N		-0,65
	TC_REG2=REGULT_Y		0,49
	TC_REG4=HEATOFF_N		-0,58
	TC_REG4=HEATOFF_Y		0,65
VS7 [-1,81 ; 2,89]	WL_EQ1	0,82	
	WL_US1	0,69	
	WL_US2	0,39	
	WL_REG1=SCREENREG_HIGH		0,59
	WL_REG1=SCREENREG_LOW		-0,84
	WL_REG1=SCREENREG_MID		0,20
VS8 [-3,3 ; 1,8]	HY_US2=BATH_N		-0,05
	HY_US2=BATH_Y		1,12
	HY_REG2=SELECSORT_N		-2,18
	HY_REG2=SELECSORT_Y		0,22
VS9 [-1,82 ; 0,85]	LI_REG1=LEDBC_ALW		-0,99
	LI_REG1=LEDBC_N		0,68
	LI_REG1=LEDBC_SMT		0,34
	LI_REG2=HALOG_N		-1,28
	LI_REG2=HALOG_Y		0,39

Tableau 28 : Barycentre des archétypes de comportement dans l'espace synthétique. Source : Auteur, calculs réalisés sur la base ENERGIHAB

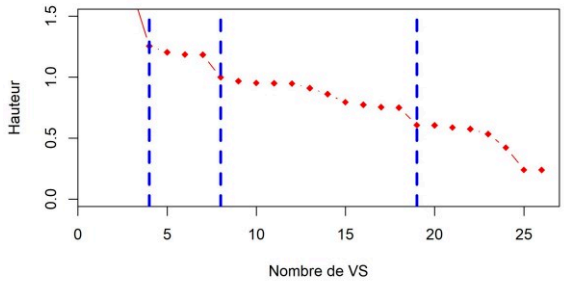
	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
Equipement alimentaire	-0,41	0,16	-0,62		0,33	0,98	
Présence moyenne au domicile	0,10	0,24	-0,53	0,63	0,34	-0,36	-0,43

Equipement et usage pour l'hygiène	-0,37	0,28	-0,70	-0,28		0,86	0,41
Comportements de restriction	-0,37		0,15	-0,37	2.72		-0,27
Demande en chauffage	0,14	-0,47	-0,20	1.04			
Régulation du chauffage	0,12	0,58	-0,30	-0,28		-0,36	-0,33
Demande en loisirs	0,60	-0,27	0,51	0,31		-1.00	-0,28
Gestes verts	0,30	0,22	0,18	0,10			-2.37
Equipement d'éclairage	-1.46	0,53	0,45	0,46	-0,57	-0,28	0,33

Annexe 2 : Classification des variables de comportement de la base PHEBUS



(A) Dendrogramme des variables de comportement. La hauteur d'un cluster C= AUB est définie comme h(C)= d(A, B)



(B) Evolution de la hauteur en fonction du nombre de clusters variables.

Figure 74: Dendrogramme (A) et Inertie (B) calculée pour la classification des variables de comportement. Source : Auteur après calculs sur la base ENERGIHAB.

Le regroupement des variables effectué résume 44% de la variance initiale. Le Tableau 29 recense les regroupements des variables ainsi que leur contribution, définie comme la corrélation au carré avec la VS. Seules les variables ayant une contribution supérieure à 20% sont étudiées. Ces variables n'ont donc pas été considérées pour l'interprétation des VS. Après avoir identifié les variables les plus fortement liées à la VS, la VS peut être interprétée en étudiant les positions des modalités des variables qualitatives sur la VS et en étudiant la corrélation des variables quantitatives avec la VS.

Tableau 29: Résumé des liens entre les VS et les variables initiales. Une interprétation des VS est proposée en accord avec les variables sélectionnées. Source : Auteur après calculs sur la base PHEBUS.

Variable synthétique	VS1	VS2	VS3	VS4	VS5	VS6	VS7	VS8
Interprétation de la variable synthétique	Niveau d' équipement	Equipement « vert »	Usage de l' ECS	Demande en chauffage	Périodes d' inoccupations longues	Présence moyenne au logement	Comportements de restriction	Régulation du chauffage
Nombre de variables dans le groupe (Degrés de liberté)	5	2	2	3	4	3	4	4
Variables inclues (corrélation entre la variable la variable synthétique VS)	F_EQ1_N (0,66) HY_EQ1_N (0,56) F_EQ2_C (0,46) WL_EQ1_N (0,39) LI_EQ1_C (0,11)	F_REG1_N (0,7) HY_REG4_N (0,7)	HY_US2_C (0,6) HY_US1_N (0,6)	TC_US1_N (0,62) TC_US4_C (0,53) TC_US5_C (0,13)	OCC2_N (0,73) TC_REG5_C (0,59) OCC5_C (0,14) TC_EQ1_C (0,01)	OCC1A_C (0,84) OCC1B_C (0,75) OCC4_N (0,38)	TC_REG7_C (0,85) TC_REG6_C (0,73) TC_REG3_C (0,6) HY_REG1_C (0,08)	TC_REG4_C (0,67) TC_REG2_C (0,5) TC_EQ2_C (0,33) TC_US2_C (0,2)