Modèle Bayésien pour la Désagrégation du chauffage

François Culière

Supervisé par Laetitia Leduc

Hello Watt

Télécom Paris

Janvier, 2020

Une dissertation rédigée pour l'obtention du diplôme du M.S Big Data Télécom Paris.

Remerciements

Je tenais à remercier Monsieur Xavier COUDERT, co-fondateur d'Hello Watt de m'avoir accueilli dans son entreprise.

Je tenais à remercier chaleureusement Monsieur Alexander BELIKOV, responsable de l'équipe data pour la qualité de son encadrement et de m'avoir permis de travailler en autonomie sur des études passionnantes et importantes pour le développement de l'entreprise.

Que ma tutrice de stage, Mme Laetitia LEDUC, data scientist chez Hello Watt, veuille bien trouver ici l'expression de ma très haute considération pour son encadrement, ses conseils et sa disponibilité.

Que M. Johan SASSI, M. Guillaume MATHERON, respectivement stagiaire et data scientist avec qui j'ai partagé le quotidien, veuillent bien trouver ici le témoignage de ma considération distinguée et l'expression de mon meilleur souvenir. La qualité de nos échanges techniques et leur alacrité ont contribué à rendre des plus agréables ce séjour studieux.

Cette page serait bien incomplète si j'omettais d'exprimer avec empressement, ma profonde reconnaissance aux enseignants de Télécom Paris pour la qualité et le niveau de leurs enseignements. Il m'ont permis sans conteste de mener à bien ce projet.

Et j'achèverai cette page en remerciant Mme Giovanna VARNI, Professeur des Universités de Télécom Paris, pour avoir accepté, le tutorat académique de ce travail.

Résumé

L'adoption des compteurs intelligents est une étape majeure de la transition énergétique et d'une gestion de l'énergie plus intelligente. Le secteur résidentiel en France représente $\approx$ 35% de la consommation d'électricité avec $\approx$ 40% (INSEE) des ménages utilisant le chauffage électrique. Le nombre de compteurs intelligents Linky déployés devrait atteindre les 35 millions en 2021. Ce rapport présente un modèle bayésien de la consommation en electricité conditionnée par la température qui permet de désagréger le chauffage de la consommation totale en electricité de manière non supervisée. Le modèle est un mélange de régressions linéaires par morceaux, caractérisé par un seuil de température, en dessous duquel il existe un mélange de deux modes pour représenter l'état latent présent/absent du ménage.

Table des matières

Table des figures

Liste des Abbreviations

NILM Non-Intrusive Load Monitoring vi
CRE commission de régulation de l'énergie 1
ADEME agence de la transition écologique 30
DPE diagnostic de performance énergétique vii
kWh kilowattheures 1
SEO Search Engine Optimisation 3
SEA Search Engine Advertising 3
HTC coefficient de transfert thermique (kW/°C) 7
HPLC coefficient de perte thermique (kW/°C) 8
MCMC Markov chain Monte Carlo 12
VI inférence variationnelle 12
ELBO borne inférieure variationnelle 13
BIC critère d'information bayésien 25
KDE estimation de densité par noyau 26
KL Kullback-Leibler 12

1

Présentation de l'entreprise

Créé fin 2016 par Sylvain Le Falher et Xavier Coudert, Hello Watt est le conseiller énergie du particulier. L'entreprise a pour but d'accompagner les ménages dans la transition énergétique, et d'aider ces derniers dans la gestion de leur consommation d'énergie à travers différents services. L'entreprise compte aujourd'hui 100 salariés, et deux services pleinement opérationnels.

1.1 | Le secteur de l'énergie en France

Pour bien comprendre l'activité de l'entreprise et percevoir les défis qui s'y rapportent, nous décrivons succinctement le fonctionnement du secteur de l'énergie en France. En 2007, les états membres de l'Union Européenne ont décidé d'ouvrir le secteur de l'énergie à la concurrence dans le cadre du marché unique permettant la libre circulation des biens, des personnes et des services en Europe. L'objectif : améliorer la compétitivité de ce secteur en rationalisant la production, le transport ainsi que la commercialisation, et ce, au bénéfice des consommateurs.

En France, les tarifs du gaz et de l'électricité sont réglementés. Ils sont fixés par la commission de régulation de l'énergie (CRE), et doivent être appliqués par EDF et Engie, fournisseurs historiques français de gaz et d'électricité. Les fournisseurs concurrents n'ont pas l'obligation, contrairement à EDF et Engie, de suivre cette tarification, et peuvent donc acheter des kilowattheures (kWh) et les revendre moins cher à leurs clients. Depuis lors, de nombreux acteurs se sont lancés dans ce secteur (Total, butagaz, Cdiscount, Ovo, Mint, Bulb...) proposant différentes offres : électricité online, électricité verte, offre éco... Il existe en France plus de 32 fournisseurs d'énergie, électricité et gaz confondus. Les offres de ces fournisseurs peuvent permettre aux particuliers de réaliser

des économies pouvant aller jusqu'à 200€ par an, avec des prix du kilowattheure jusqu'à 10 ou 15% inférieurs au tarif réglementé.

1.2 | Hello Watt

1.2.1 | Les services proposés

Pour clarifier les offres proposées par les fournisseurs d'électricité et de gaz, Hello Watt a mis en place un comparateur permettant aux particuliers de trouver en quelques clics le fournisseur d'énergie correspondant au mieux à leurs besoins. L'entreprise s'inscrit comme un intermédiaire indépendant entre les particuliers et les fournisseurs et donne des avis et conseils impartiaux concernant les différentes offres d'électricité et de gaz présentes sur le marché. Une équipe de conseillers est disponible par téléphone, pour orienter chaque particulier et réaliser les démarches administratives associées au changement de fournisseurs.

Cette activité, bien qu'aujourd'hui fructueuse au vu de l'état du marché Français, ne semble pas être pérenne et n'apporte qu'une valeur relative. L'ambition d'Hello Watt va bien au delà. L'objectif est en effet d'accompagner le client à chaque étape de sa transition énergétique quelle qu'elle soit. On distingue trois axes de développement, correspondant à trois grands stades de la transition énergétique des particuliers (Fig. 1.1) :

FIGURE 1.1 – Axes de développement de Hello Watt

Hello Watt a crée deux nouveaux départements pour répondre à son ambition : les départements solaire et rénovation énergétique. Ce dernier à pour mission d'accompa-

gner les particuliers dans la rénovation énergétique de leur logement (installation de chaudière, isolation des fenêtres...) par la mise en relation avec les installateurs, et en favorisant l'accès aux différentes aides et primes mises en place par le gouvernement pour accélérer et compléter la transition énergétique des particuliers. L'équipe data d'Hello Watt est au cœur du développement de ce département. Nous détaillerons plus précisément les enjeux en section 2.3.

1.2.2 | L'organisation de l'entreprise

L'entreprise peut être divisée en 4 grands départements, contenant plusieurs équipes.

Le pôle opération

- **Les conseillers**, experts en énergie qui ont pour mission de répondre aux demandes clients par téléphone et les orienter vers le choix du fournisseur d'énergie le plus adapté à leurs besoins.
- **Les business analysts** suivent et gèrent les performances des commerciaux et proposent des axes d'améliorations au vu de l'évolution du marché.
- **L'équipe formation** forme les conseillers et autres collaborateurs aux enjeux et évolutions du domaine de l'énergie via des formations organisées occasionnellement.

Le pôle marketing

- **L'équipe Search Engine Optimisation (SEO)**, responsable du référencement naturel du site sur les moteurs de recherche pour optimiser l'acquisition naturel (SEO) et payante (Search Engine Advertising (SEA)) de clients.
- **L'équipe contenu** rédige les articles du site Hello Watt pour lui donner de la visibilité sur les sites de référencement.

Le pôle technique

- **Les développeurs**, chargés de maintenir le site de d'implémenter de nouvelles fonctionnalités afin d'enrichir l'expérience utilisateur.
- **L'équipe data** gère tout ce qui est relatif au traitement et à l'exploitation de données.

Le pôle business développement

- **L'équipe solaire** développe et crée des partenariat avec des installateurs solaires.

— **L'équipe rénovation** est chargée du développement et de la mise en place des futurs services d'Hello Watt concernant la rénovations.
— **L'équipe partenariat** contribue à trouver de nouveaux partenaires chez les fournisseurs, installateurs solaires etc... et gère les relations avec les partenaires historiques.

1.3 | Le pôle Data Science

Ce pôle joue un rôle fondamental dans les perspectives d'évolution d'Hello Watt. Il répond à 3 grands besoins :

1. Développer les modèles pour le Coach Conso. Celui-ce a une place stratégique dans le développement d'Hello Watt. C'est en effet un concentré de la valeur ajoutée d'Hello Watt où nous pouvons informer et aider le client de manière personalisée grâce à ses données de consommation. Tout le travail effectué sur les données du client l'est par notre équipe, avec l'aide de quelques développeurs back-end pour le traitement et le stockage de la donnée, ainsi que pour la mise en production des algorithmes. Toute estimation proposée dans le Coach Conso est le résultat d'un modèle développé par l'équipe data, entraîné sur toutes les données que l'on a à disposition.
2. Optimiser les dépenses marketing. Hello Watt a un fort besoin de trouver de nouveaux prospects en permanence. Ainsi, une grande partie des dépenses opérationnelles de Hello Watt va dans l'acquisition de nouveaux clients au travers du SEA, notamment via Google Adwords. Toutes les données d'impression, de clic et de conversion étant stockées, il est possible de les analyser pour optimiser la stratégie SEA.
3. Répondre à d'autres besoins ponctuels utilisant des données. Cela inclut par exemple l'extraction d'informations sur des factures d'électricité, la mise à disposition de bases de données croisées pour les développeurs afin d'ajouter un graphique sur le site ou des statistiques pour l'équipe contenu.

2

Contexte général & Objectifs

La réduction de la consommation d'énergie est un défis écologique clé du 21ème siècle. Beaucoup d'attention a été donnée au secteur résidentiel dans la communauté de la désagrégation d'énergie ([1, 2, 3, 4]), et en particulier à la détection d'anomalies liées aux appareils électroménagers ([5, 6]) ou à l'isolation des maisons ([7, 8, 9, 10, 11]). En France, 30% de l'énergie est consommée par le secteur résidentiel. Cela représente environ 37 [1] milliards d'euros par an. Il existe pourtant un potentiel d'économies important. Mais pour que chaque particulier puisse agir et participer à la transition énergétique, il lui faut comprendre comment évolue la consommation d'énergie de son foyer et quels sont les bons gestes à adopter en conséquence. C'est ici que l'analyse des données de consommation entre en jeu, en permettant aux particuliers de comprendre leur façon de consommer et en leur proposant des recommandations personnalisées ceux-ci pourront directement agir pour réaliser des économies. Dans le contexte de la régulation européenne de l'energie, les compteurs électriques Linky et de gaz Gazpar sont largement déployés. On attends à ce que leurs nombre atteigne 35M en 2021 et 11M en 2022 respectivement. Aujourd'hui, plus de 2/3 des foyers en France sont équipés de compteurs Linky. Cette quantité de données détaillées disponibles (avec l'accord des particuliers) change la donne et rend désormais possible ces analyses.

2.1 | Désagrégation d'énergie

La désagrégation d'énergie consiste à analyser les signaux de capteurs de courant et de tension électriques provenant de plusieurs appareils pour en déduire leurs réparti-

1. INSEE, Institut National de la Statistique et des Etudes Economiques, disponible sur https://www.insee.fr/fr/accueil (28.5 millions de ménages consommant en moyenne 1300€ pour l'énergie de leur logement (consulté le 03/12/2020))

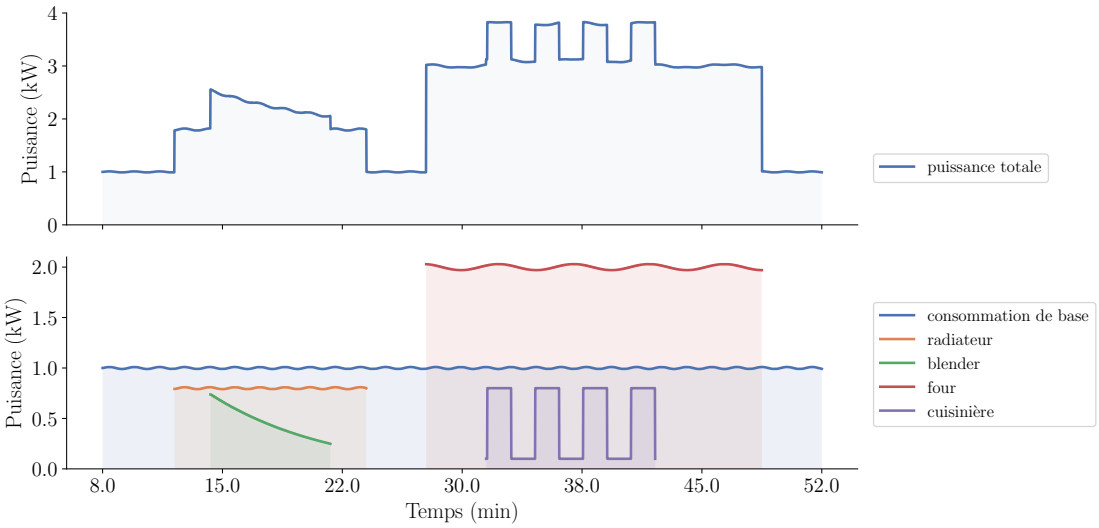

FIGURE 2.1 – Illustration de la désagrégation d'énergie

tions sur une période donnée (Fig. 2.1).

2.1.1 | Les méthodes

Les méthodes employées pour la désagrégation d'énergie dépendent fortement du type de mesure à disposition. Les mesures peuvent provenir d'un capteur mesurant la puissance agrégée où de plusieurs mesurant des groupes d'appareils. La première solution à l'avantage de ne pas nécessiter d'installation supplémentaire pour les particuliers lorsqu'ils possèdent un compteur communiquant de type Linky ou Gazpar. Le type de mesure se différencie aussi par leurs périodes d'échantillonnage pouvant aller de quelques millisecondes à 30 minutes. Le problème de la désagrégation est d'autant plus difficile que le nombre d'appareils mesurés par capteur est grand est que la fréquence d'échantillonnage est faible. Chez Hello Watt, nous utilisons les mesures basses fréquences des compteurs communiquants. Dans ce contexte, nous nous concentrons sur les appareils ayant le plus d'impact sur la consommation totale des logements comme le chauffe-eau, le chauffage, ou le four.

2.1.2 | La communauté NILM

La communauté NILM est un groupe de chercheurs académiques et industriels qui travaillent sur la désagrégation d'énergie. En 2020, Hello Watt à participé à la 5ème confé-

rence de la communauté comptant plus de 250 participants. L'évolution du nombre de participants aux conférences et workshops de la communauté NILM montre l'intérêt croissant porté par la communauté scientifique à ce sujet.

2.2 | Désagrégation du chauffage

Nous nous intéressons dans ce rapport plus particulièrement à la désagrégation de la part du chauffage dans la consommation totale. Grâce à la relation directe entre la température extérieure et les pertes de chaleur des logements ([1, 10, 11]), la séparation entre la consommation totale du logement de la partie de chauffage peut être réalisée de façon non supervisée. Cette approche est applicable à grande échelle en France où la consommation de chauffage compte pour 40% [2] de la consommation annuelle d'électricité. En plus de la désagrégation du chauffage, le modèle présenté en section 3.4 permet d'évaluer la qualité d'isolation d'un logement en comparant son coefficient de transfert thermique (kW/°C) (HTC) avec des logements de surface et année de construction similaires.

2.2.1 | Lien entre la puissance totale mesurée et la température extérieure

Nous présentons dans cette section la relation entre la puissance totale mesurée d'un logement et la température extérieure. La relation présentée justifie l'utilisation d'un modèle non supervisé pour le calcul du chauffage et pour l'évaluation de la qualité d'isolation des logements. Par soucis de concision, les noms des variables se veulent le plus explicite possible et ne seront pas tous détaillés. Le lecteur pourra se référer à la référence [11] pour de plus amples détails.

L'équation d'équilibre des flux thermiques d'un logement peut s'écrire :

$$\Phi_{tot} = \Phi_{\text{conduction}} + \Phi_{\text{ventilation}} + \Phi_{\text{solaire}} + \Phi_{\text{appareils}} + \Phi_{\text{occupants}}$$

En linéarisant les apports solaires et les pertes de chaleurs par ventilation, la puissance totale mesurée du logement s'écrit :

$$P_{tot} = \text{HPLC}\,(T_{in} - T_{ext}) - \frac{1}{\eta_{HS}} A_{sol} I_{sol} - \frac{\eta_B}{\eta_{HS}} P_B - \frac{\Phi_O}{\eta_{HS}} + P_B \tag{2.1}$$

2. À partir d'une étude réalisée par l'INSEE en 2016.

où

P_{tot} = puissance totale mesurée du logement

η_{HS} = rendement des installations de chauffage

A_{sol} = surface équivalente du logement absorbant les radiations

I_{sol} = irradiance de la surface du logement

η_B = fraction de la consommation de base participant au chauffage (chauffe eau...)

P_B = consommation de base du logement

$\text{HPLC} = \dfrac{\text{HTC}}{\eta_{HS}}$, le coefficent de perte thermique

$$\text{HTC} = \frac{d\left(\Phi_{\text{conduction}} + \Phi_{\text{ventilation}}\right)}{d\Delta T} = H_{\text{conduction}} + \frac{d\Phi_{\text{ventilation}}}{d\Delta T}$$

L'équation précédente dépend de la température intérieure qui restera inconnue dans le modèle. En faisant une hypothèse de relation linéaire entre le température intérieure et extérieure on peut relier la variation de la puissance mesurée avec la température extérieure et le coefficient de perte thermique (kW/°C) (HPLC) en ajoutant un coefficient adimensionnel F_T qui représente la variation de T_{in} avec T_{ext}. On trouve alors :

$$\frac{dP_{tot}}{dT_{ex}} = -\text{HPLC}(1 - F_T)$$

Ces résultats montrent que l'on peut trouver le HPLC en utilisant seulement les données de consommation des compteurs communiquant et des données météorologiques. Ce coefficient permet aussi de désagréger le chauffage de la consommation totale en utilisant la relation :

$$P_{\text{chauffage}} = \Phi_{\text{conduction}} + \Phi_{\text{ventilation}} = \text{HPLC}(T_{ext} - T_{in}) \quad \text{avec } T_{in} = f(T_{ext}, F_T)$$

Dans la pratique, le coefficent F_T est supposé être le même pour toutes les maisons même si on s'attend à avoir une dépendance de ce coefficient avec la température moyenne extérieure annuelle du logement. En effet, on peut supposer que les habitants des zones chaudes tolèrent moins le froid que celles des régions plus froides. De plus, le rendement des installations de chauffage η_{HS} est inconnu. On le supposera aussi constant pour tous les logements. Cette dernière hypothèse regroupe les anomalies d'isolation et de rendement des installations de chauffage. Il faudra veiller à ce détail lors de l'analyse de l'isolation des logements.

Ainsi on confondra HPLC, $\text{HPLC}(1 - F_T)$ et $\text{HTC} = \eta_{HS}\text{HPLC}$ pour l'évaluation de l'isolation des maisons. T_{in} sera supposée fixe et inférée par l'algorithme comme le point de cassure à partir duquel on observe une hausse significative de la consommation d'électricité (voir 3.4).

2.3 | Analyse de la consommation chez Hello Watt

2.3.1 | Présentation du coach conso

C'est un service développé par Hello Watt permettant aux particuliers de suivre leur consommation électrique et de bénéficier de conseils ou d'alertes sur celle çi. Il est disponible directement sur le site internet d'Hello Watt. Sur le Coach Conso, le client peut visualiser sa consommation, voir des prédictions de consommation personnalisées, estimer la qualité de l'isolation de son logement ou encore accéder à une estimation de la désagrégation de sa consommation par type d'usage ou d'appareil. Tous ces estimateurs sont le résultat de modèles développés par le pôle data. Ce service vise à fidéliser le client en l'aidant à suivre sa consommation ; il est activement développé en ce moment.

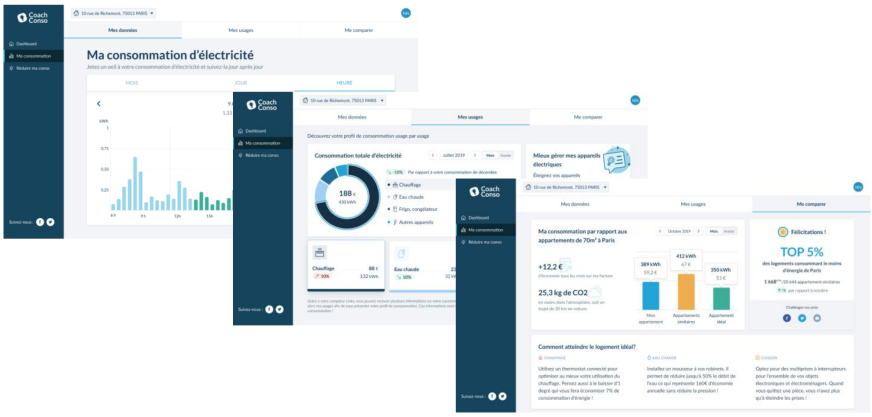

FIGURE 2.2 – Coach conso

2.3.2 | Les enjeux de l'activité rénovation

L'équipe chargée de l'activité de rénovation est très récente chez Hello Watt mais représente un axe stratégique important. En effet, les enjeux environnementaux ont poussé le gouvernement à encourager fortement les rénovations énergétiques, avec un objectif de 500 000 rénovations globales d'ici 2030, ce qui correspond à un marché de 14Md€ sur 10 ans. Pour cela, beaucoup d'aides sont proposées. Cette action, liée à la prise de conscience des particuliers de plus en plus soucieux de l'environnement et de l'impact de leur consommation sur leur finances, devrait se traduire en une forte croissance du marché. Cependant, il existe beaucoup de frictions dans ce domaine, résumées dans la figure 2.3.

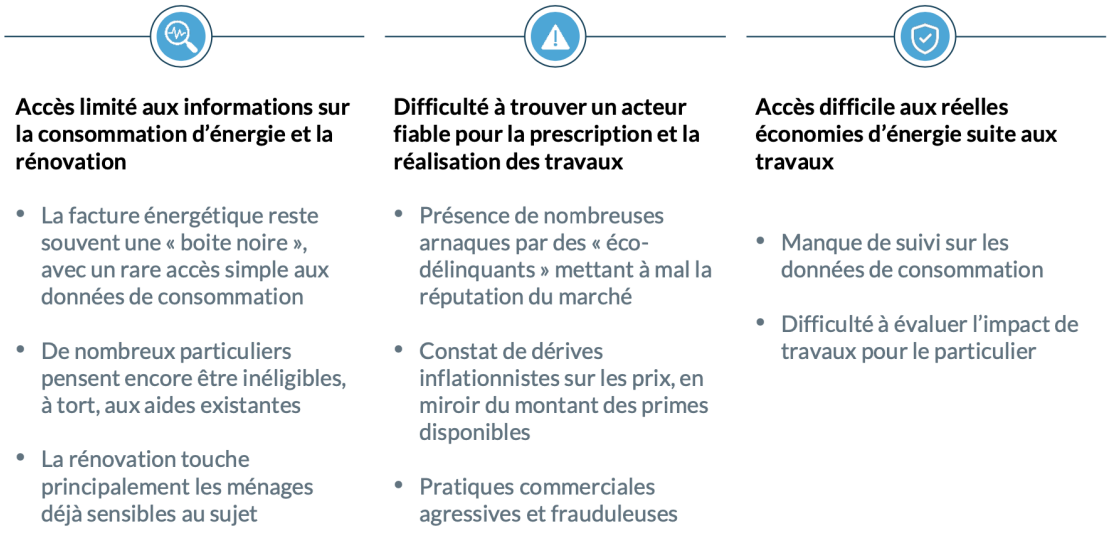

FIGURE 2.3 – Maturité du marché de la rénovation

Le Coach Conso a également pour vocation de résoudre en partie le premier et le dernier problème. En analysant les données de consommation et la situation des clients, Hello Watt est capable d'informer les clients sur l'intérêt qu'ils ont ou non à effectuer des rénovations via cette interface. Une fois les rénovations effectuées, des modèles peuvent estimer l'impact des travaux et les gains réels. Hello Watt fait parti des lauréats du hackathon RenovAction [3] avec Mon Coach Renov comme prototype de cette solution. L'équipe chargée du développement de l'activité rénovation, en plus de participer au développement de ces aspects du Coach Conso, s'attaque également au deuxième problème, en cherchant par exemple des partenariats.

3. https://www.hackathon-renovaction.fr/program/hackathon, organisé par le ministère de la transition écologique et solidaire et le ministère de la cohésion des territoires et des relations avec les collectivités territoriales

3

Méthodes & Outils

Le modèle de désagrégation du chauffage proposé est un modèle bayésien implémenté avec la librairie de programmation probabiliste Pyro. Cette partie vise à présenter les outils et méthodes nécessaires au développement du modèle.

3.1 | Inférence bayésienne

Le bayésianisme est une forme d'épistémologie qui connaît un succès croissant dans de nombreux domaines du savoir. Avant 1990, ce paradigme est délaissé en faveur du fréquentisme pour deux causes racines : l'opposition philosophique à définir les probabilités de façon subjectives et le manque de ressources informatiques pour réaliser les calculs nécessaires aux analyses bayésiennes [12, 13].

De façon formelle, le paradigme bayésien peut se résumer par la formule suivante :

$$\begin{aligned} \mathrm{P}(\Theta \mid \text{data}) &= \frac{\mathrm{P}(\text{data} \mid \Theta) \times \mathrm{P}(\Theta)}{\mathrm{P}(\text{data})}, \\ \text{Posterior} &= \frac{\text{Vraisemblance} \times \text{Prior}}{\text{Evidence}}. \end{aligned}$$

Cette formule permet de comprendre les causes d'opposition de ce paradigme dans laquelle on retrouve une définition d'une probabilité à priori ($\mathrm{P}(\Theta) = \text{Prior}$) qui défini de façon subjective la connaissance que l'on a sur une hypothèse et le calcul de l'évidence ($\mathrm{P}(\text{data}) = \int \mathrm{P}(\text{data} \mid \Theta)\mathrm{P}(\Theta) d\Theta$) souvent non calculable puisqu'elle requiert un calcul d'intégral sur un espace possiblement très grand (imaginons les poids d'un réseau de neurones ou un problème de mixture où la combinatoire explose).

Notons que la critique de subjectivité de la pensée bayésienne peut être défendue par deux arguments : (1) les distributions à priori peuvent être rendues aussi dispersées que

souhaité, (2) elles peuvent provenir d'études qui ne se basent pas sur la seule subjectivité du scientifique. Ceci peut être illustré avec l'estimation de la moyenne d'une Gaussienne d'écart type connu (Annexe A). Les calculs montrent que par rapport à l'estimateur de maximum de vraisemblance, l'estimateur bayésien est biaisé mais peut avoir une variance plus faible lorsque l'information à priori est précise.

Quand au dernier point, les méthodes de Markov chain Monte Carlo (MCMC) et d'inférence variationnelle (VI) permettent de contourner l'impossibilité du calcul en estimant les distributions a posteriori. Le méthode MCMC approche la distribution en construisant une marche aléatoire sur la chaîne de Markov ayant pour loi stationnaire la distribution a posteiori. La VI utilise une distribution variationnelle paramétrée approchant la distribution a posteriori au sens de la divergence de Kullback-Leibler (KL). Nous allons détailler un peu plus la deuxième méthode puisque c'est la méthode que nous avons retenue pour l'implémentation de notre modèle de désagrégation du chauffage

Nous verrons section 3.4 que notre modèle peut s'écrire comme un réseau bayésien ayant pour distribution jointe :

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}),$$

où

$x = (\text{Température}, \text{Consommation})$

z correspond aux variables aléatoires du réseau bayésien.

Il comprend les pentes, biais et points de cassure du modèle ainsi que l'espace latent $\in (0, 1)$ qui modélise la présence où non des habitants dans le logement.

θ les paramètres des distributions utilisées pour modéliser les variables aléatoires.

3.2 | Inférence variationnelle

On se place dans le contexte où nous modélisons le processus génératif des données avec un modèle probabiliste ayant des variables latentes z et des paramètres θ. Le modèle à une probabilité jointe de la forme :

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}).$$

Le but de tout algorithme d'inférence est d'apprendre les paramètres θ du modèle qui maximisent la log évidence des données. C'est à dire trouver les paramètres θ donnés par

$$\theta_{\max} = \underset{\theta}{\operatorname{argmax}} \log p_\theta(\mathbf{x}),$$

où la log évidence $\log p_\theta(\mathbf{x})$ est définie par :

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z}.$$

Ce problème est doublement difficile puisque pour θ fixé, l'intégrale sur les variables latentes z est souvent non calculable et dans le cas où l'on sait calculer cette intégrale pour toute les valeurs de θ, la maximisation de la log évidence comme une fonction de θ est un problème d'optimisation non convexe difficile.

En plus de trouver $\theta_{\max}$, on aimerait calculer la distribution a posteriori des variables latentes sachant les données :

$$p_{\theta_{\max}}(\mathbf{z} \mid \mathbf{x}) = \frac{p_{\theta_{\max}}(\mathbf{x}, \mathbf{z})}{\int p_{\theta_{\max}}(\mathbf{x}, \mathbf{z}) d\mathbf{z}}.$$

Comme détaillé précédemment, la log évidence qui est au dénominateur de cette expression n'est pas calculable. La VI est une solution pour trouver le paramètres $\theta_{\max}$ et trouver une approximation de la distribution a posteriori $p_{\theta_{\max}}(\mathbf{z} \mid \mathbf{x})$. L'idée est d'introduire une distribution paramétrée $q_\phi(z)$ appelé distribution variationnelle qui va approcher la distribution $p_{\theta_{\max}}(\mathbf{z} \mid \mathbf{x})$ au sens de la divergence de Kullback-Leibler.

On souhaite maintenant trouver les paramètres ϕ, θ donnés par :

$$\phi_{\max}, \theta_{\max} = \underset{\phi,\theta}{\operatorname{argmin}} \mathrm{KL}(q_\phi(z) \| p_\theta(z \mid \mathbf{x})).$$

Cette dernière expression n'est pas calculable puisqu'elle dépend directement de la probabilité a posteriori $p_\theta(\mathbf{z} \mid \mathbf{x})$. Pour résoudre ce problème, on considère la propriété suivante qui met en évidence la quantité qui sera maximisée, la borne inférieure variationnelle (ELBO) définie par :

$$\mathrm{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z})} \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}) \right].$$

L'ELBO est une borne inférieure de la log évidence, pour tout θ et ϕ, on a :

$$\log p_\theta(\mathbf{x}) \geq \mathrm{ELBO}.$$

Ainsi, en maximisant cette quantité, on augmente la log évidence des données. On peut montrer que la différence entre la log évidence et l'ELBO est donnée par la divergence KL entre la distribution variationnelle et la distribution a posteriori :

$$\log p_\theta(\mathbf{x}) - \text{ELBO} = \text{KL}(q_\phi(z) \| p_\theta(z \mid \mathbf{x})).$$

Notons aussi que l'ELBO peut s'écrire aussi comme :

$$\text{ELBO} = H(q_\phi) - H(q_\phi, p_\theta) \quad \text{où H désigne l'entropie et l'entropie croisée.}$$

Ainsi, en maximisant l'ELBO, on souhaite en même temps maximiser l'entropie de q_ϕ tout en minimisant l'entropie croisée $H(q_\phi, p_\theta)$. Cela montre aussi que la distribution variationnelle cherche à expliquer au mieux les données tout en ayant une distribution pas trop étroite pour les expliquer.

3.3 | Programmation probabiliste

Nous nous intéressons maintenant au language permettant de réaliser l'inférence des modèles probabilistes. Il existe plusieurs librairies permettant de faire de la programmation probabiliste. Nous citons ici PyMC3 basée sur Theano , Edward sur Tensorflow et enfin Pyro sur Pytorch.

Dans le cadre de l'inférence variationnelle, ces librairies ont un objectif commun :

1. Décrire un modèle probabiliste,
2. Décrire une distribution variationnelle,
3. Minimiser l'ELBO , nécessitant le calcul et la rétro-propagation de son gradient.

La dernière étape est la raison pour laquelle ces librairies se servent des environnements de calcul différentiel comme moteur de calcul. L'optimisation peut ainsi être effectuée pour des jeu de données de grandes tailles en un temps respectable avec l'utilisation des GPU et du batching.

Calcul du gradient L'apprentissage des paramètres du modèle passe par le calcul du gradient de l'ELBO :

$$\nabla_{\theta,\phi}\text{ELBO} = \nabla_{\theta,\phi}\mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z})\right].$$

La difficulté du calcul réside dans le caractère stochastique de l'expression et dans la dépendance de l'espérance mathématique $\mathbb{E}_{q_\phi}$ avec les paramètres ϕ à optimiser.

Supposons que l'on puisse reparamétrer l'expression de la façon suivante (c'est le cas pour une distribution normale) :

$$\mathbb{E}_{q_\phi(\mathbf{z})}\left[f_\phi(\mathbf{z})\right] = \mathbb{E}_{q(\epsilon)}\left[f_\phi\left(g_\phi(\epsilon)\right)\right].$$

On peut alors déplacer l'opérateur ∇ dans l'espérance mathématique :

$$\nabla_\phi \mathbb{E}_{q(\epsilon)}\left[f_\phi\left(g_\phi(\epsilon)\right)\right] = \mathbb{E}_{q(\epsilon)}\left[\nabla_\phi f_\phi\left(g_\phi(\epsilon)\right)\right],$$

et obtenir une estimation non biaisée du gradient par la méthode de Monte Carlo avec des échantillons de cette espérance mathématique.

Dans le cas où cette reparamétrisation est impossible (c'est le cas pour la plupart des distributions), le gradient peut être calculé en utilisant l'identité :

$$\nabla_\phi q_\phi(\mathbf{z}) = q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}),$$

ce qui permet de réécrire la quantité d'intérêt de la façon suivante et appliquer de nouveau la méthode de Monte Carlo :

$$\mathbb{E}_{q_\phi(\mathbf{z})}\left[\left(\nabla_\phi \log q_\phi(\mathbf{z})\right) f_\phi(\mathbf{z}) + \nabla_\phi f_\phi(\mathbf{z})\right].$$

Nous voyons ici un des problème de la méthode d'inférence variationnelle. Bien que les gradients estimés soit sans biais, ils vont avoir une grande variance à cause de leur estimation par méthode de Monte Carlo. Cette variance est bien visible lorsque l'on trace l'ELBO en fonction du nombre de pas de gradient effectués dans le programme d'estimation.

Des méthodes connues sous le nom de Rao-Blackwellization permettent de limiter la variance de l'estimation du gradient.

3.4 | Modèle de désagrégation du chauffage

3.4.1 | Description

Nous présentons dans cette partie le modèle de désagrégation non supervisé du chauffage fondé sur le modèle physique des pertes thermiques d'un bâtiment de la section 2.2. Le modèle développé est un modèle bayésien qui décrit une mixture de régressions par morceaux dans le plan (température, consommation). Les deux modes supposés de consommation correspondent aux moments où les habitants sont présents et absents de chez eux. Le modèle est illustré figure 3.1. Ce graphique permet de voir la relation linéaire entre la température extérieure et la consommation électrique lorsque les habitants sont chez eux en dessous de la température critique présentée avec l'équation 2.1.

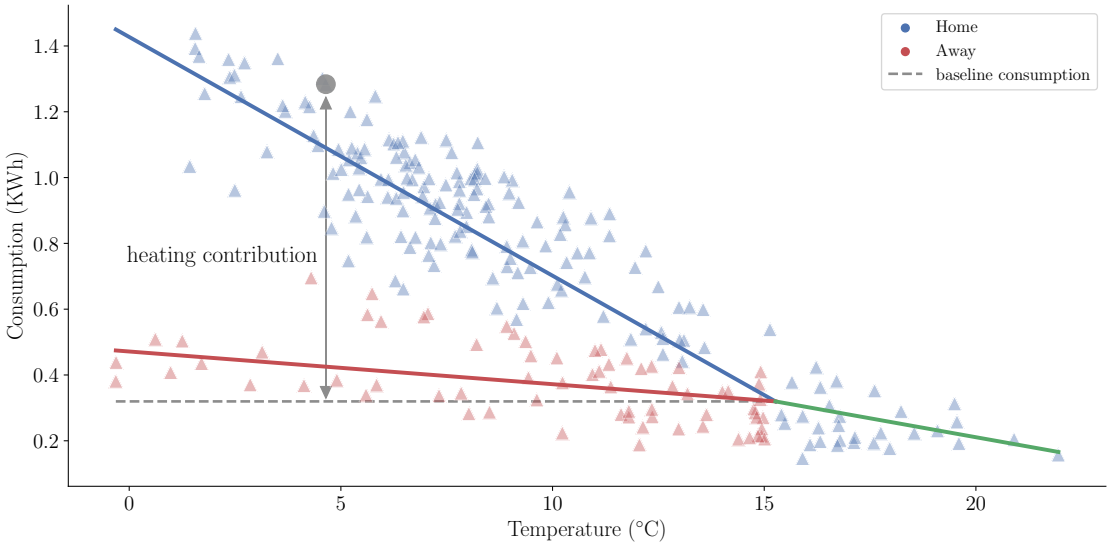

FIGURE 3.1 – Présentation du modèle dans le plan (température, consommation)

3.4.2 | Inférence

Les observations correspondent aux tuples (c_i, T_i), où c_i et T_i sont respectivement les données de consommation normalisées par la surface du logement et de température moyenne de la journée. Pour chaque logement ajusté avec le modèle on applique avant l'entraînement les transformations suivantes : $T \leftarrow \frac{T}{T_{scale}}$; $c \leftarrow \frac{c}{\text{surface} \cdot c_{scale}}$.

De cette manière, les variables d'intérêt du modèle deviennent de l'ordre de grandeur de 1, facilitant la convergence du modèle graphique. Nous exposons ici les grandes lignes du modèle bayésien proposé :

1. Choisir une température de cassure $T_k \sim \tau(Dir(\alpha_T))$.
2. Choisir une ordonnée à l'origine $b \sim \mathcal{N}(b_{loc}, b_{scale})$.
3. Choisir les pentes à droite de la cassure $w_R \sim \mathcal{N}(w_{R,loc}, w_{R,scale})$.
4. Choisir les pentes à gauche de la cassure $w_m \sim \tau(Dir(\alpha_s))$.
5. Calculer les ordonnées à l'origine restante $b_{m+1} = (w_m - w_{m+1})T_k + b_m$ (par continuité).
6. Choisir les poids de la mixture $\omega \sim Dir(\alpha_\omega)$.
7. Pour chaque tuple (c_i, T_i) : si $T_i < T_k$ échantillonner $z_i \sim Categorical(\omega)$ puis $\tilde{c_i} \sim \mathcal{N}(w_{z_i} T_i + b_{z_i}, \sigma_{z_i})$, sinon $\tilde{c_i} \sim \mathcal{N}(w_R T_i + b_R, \sigma_R)$.

La transformation τ est une combinaison de la somme cumulative de la transformation sigmoïde inverse utilisée pour échantillonner un vecteur aléatoire ordonné. Ainsi,

tout échantillon de la distribution $T_k \sim \tau(Dir(\alpha_T))$ renvoie vers un vecteur $x = [x_1, ..., x_n]^T \in \mathcal{R}^n$ tel que $x_1 < x_2 < ... < x_n$. La variable d'état latent z_i capture les états de présence et absence des habitants. Le modèle graphique est représenté figure 3.2.

Les distributions a posteriori et les meilleurs ajustements pour deux logements sont présentés figure 3.3-3.4. Nous utilisons pour toutes les données de consommation des utilisateurs du coach conso d'Hello Watt les mêmes distributions a priori $w_{R,loc}$, $w_{R,scale}$, b_{loc}, b_{scale}, α_T, α_s, α_ω.

On note que le modèle décrit ci-dessus n'est qu'un exemple, et le formalisme admet de fixer un nombre arbitraire de seuils de température T_k ainsi qu'un nombre arbitraire de composants du mélange. Cette généralisation est potentiellement utile dans le cas où les modes de consommation sont différents entre les régimes froid, intermédiaire et chaud, ou lorsque nous pouvons identifier plus de deux états latents, par exemple une famille de deux recevant deux invités pendant une semaine. Pour l'usage actuel et dans ce qui suit, nous ne considérons qu'un seul seuil de température critique T_c.

3.4.3 | Désagrégation du chauffage

Pour chaque maison ajustée avec le modèle, nous estimons la fraction de consommation due au chauffage $c^{(h)}$: pour un tuple donné $(c^{(tot)}, T)$ on évalue l'espérance mathématique $E\left[c^{(h)} \mid c^{(tot)}\right]$ et la variance $Var\left[c^{(h)} \mid c^{(tot)}\right]$ de la consommation due au chauffage.

Nous supposons que la consommation pour les températures inférieures à T_c est la somme de la consommation de base $c(T_c) = w_a T_c + b_a$ (supposée être indépendante de la température actuelle et estimée comme la consommation à $T = T_c$) et du chauffage $c^{(h)}$, qui constituent ensemble $c^{(tot)}$ (voir figure 3.1). On écrit donc :

$$c^{(h)} \mid c^{(tot)}, w_a, T_c, b_a = c^{(tot)} - w_a T_c - b_a.$$

Nous décrivons ensuite la distribution de $c^{(h)}$ en utilisant les distributions a posteriori inférées w_a, T_c et b_a , approximées par des distributions normales $w_a \sim \mathcal{N}(\mu_{w_a}, \sigma_{b_a})$, $T_c \sim \mathcal{N}(\mu_{T_c}, \sigma_{T_c})$, $b_z \sim \mathcal{N}(\mu_{b_z}, \sigma_{b_z})$. On obtient alors :

$$\begin{aligned} E\left[c^{(h)}\right] &= c^{(tot)} - \mu_{w_a}\mu_{T_c} - \mu_{b_a}\,, \\ Var\left[c^{(h)}\right] &= \left(\sigma^2_{w_a} + \mu^2_{w_a}\right)\left(\sigma^2_{T_c} + \mu^2_{T_c}\right) - \mu^2_{w_a}\mu^2_{T_c} + \sigma^2_{b_a}\,. \end{aligned}$$

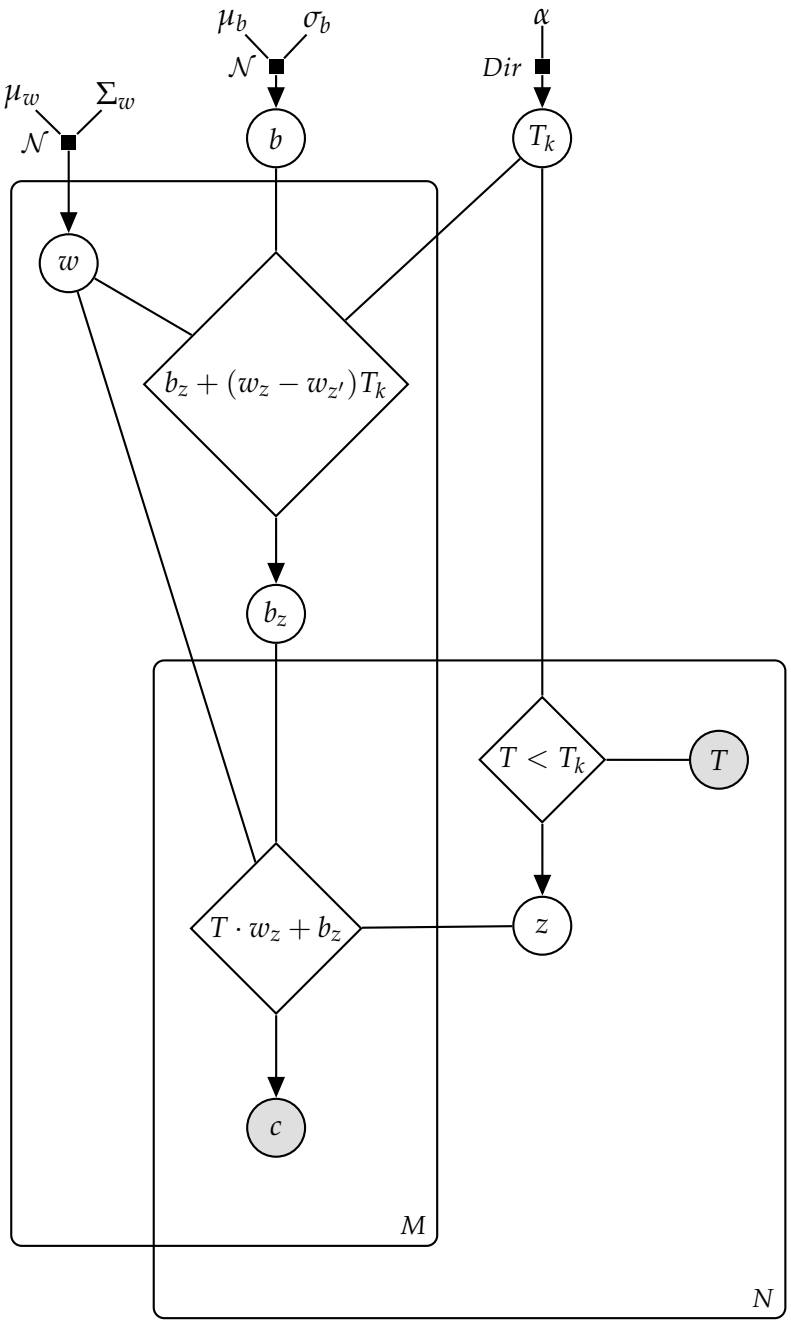

FIGURE 3.2 – Représentation graphique du modèle de mixture de régression par morceau. Le plateau à gauche représente les poids de la mixture. Celui de droite représente les observations individuelles.

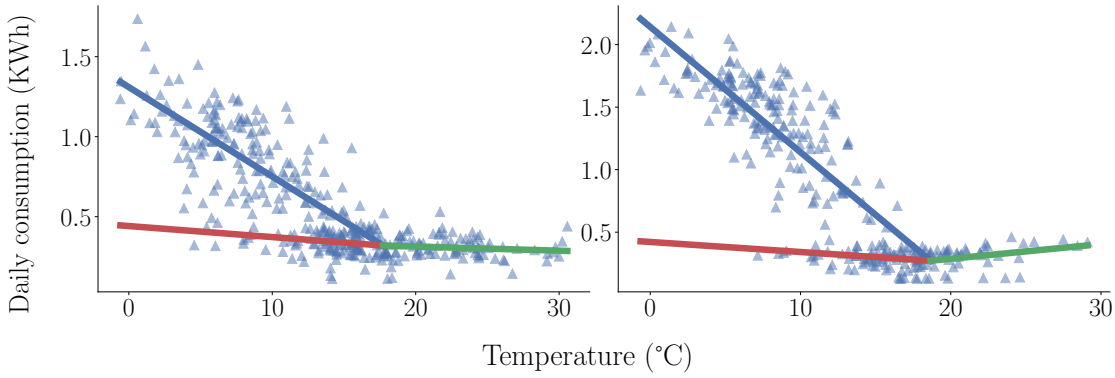

FIGURE 3.3 – Exemple de ménages avec un comportement de consommation bi-modal avec deux états présent/absent en dessous d'une température critique. Les données sont ajustées avec le modèle de mixture de régressions linéaires par morceaux.

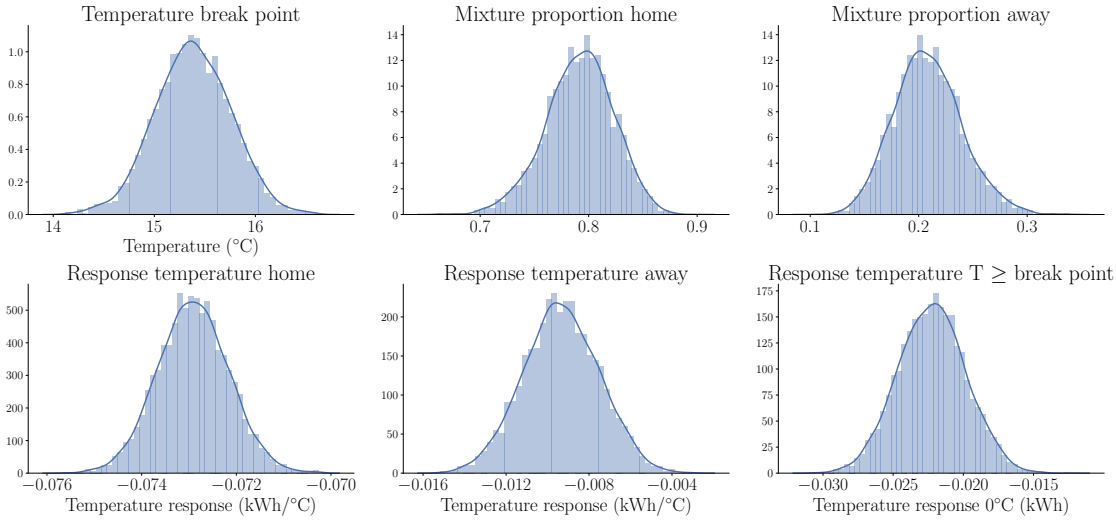

FIGURE 3.4 – Distributions a posteriori $p_\theta(z \mid \mathbf{x})$

FIGURE 3.5 – Exemple de désagrégation du chauffage pour un ménage. En haut : consommation et température en fonction du temps, en bas : consommation et température en fonction du temps avec application d'une moyenne mobile de sept jours.

Notre approche nous permet également de déduire l'état latent z en maximisant la probabilité que la consommation observée appartienne à l'état présent/absent :

$$z^* = \underset{z \text{ in } home,away}{\text{argmax}} \; P(c^{(tot)} | b_z, w_z, \sigma_z, T_c).$$

Alors que ni la température actuelle ni l'état présent/absent n'entrent dans l'estimation de la fraction de chauffage, la division de la consommation en dessous de T_c en deux états améliore la qualité globale de l'ajustement et donc l'inférence de la consommation de chauffage également. La figure 3.5 présente un exemple de désagrégation du chauffage.

4

Résultats & Discussion

4.1 | Validation du modèle

Pour valider l'approche de la désagrégation du chauffage, nous partitionnons l'ensemble des observations de consommation pour chaque logement j en deux sous-ensembles A et B, avec des observations respectivement avant et après la mi-janvier (figure 4.1).

Nous considérons comme vérité terrain la prédiction $c^{(h,AB)}$ de la partie chauffage en ajustant le modèle sur $A \bigcup B$ et en calculant les prédictions de chauffage sur B, et comme prédiction du modèle, nous considérons la contribution au chauffage $c^{(h,A)}$ estimée sur B avec le modèle ajusté sur A. Pour quantifier la qualité de la prédiction, nous calculons l'erreur quadratique moyenne relative (RMSE) :

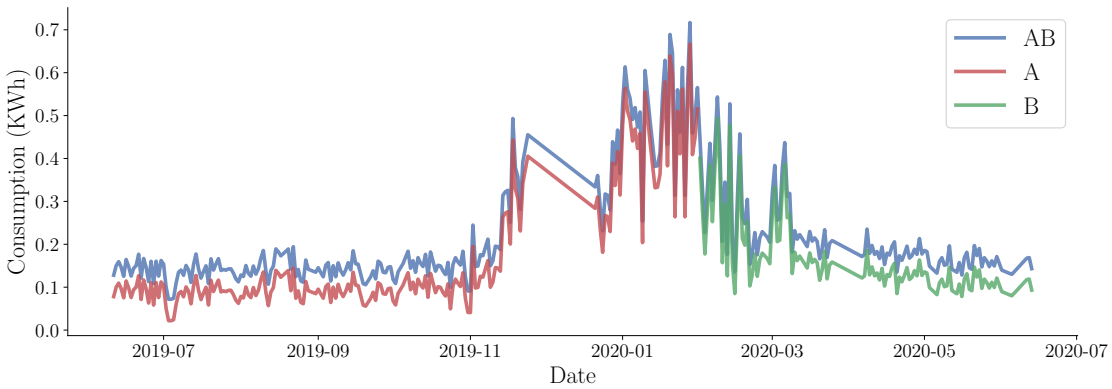

FIGURE 4.1 – Découpage des données pour valider le modèle de chauffage

$$\delta_j = \sqrt{\frac{1}{n_j} \sum_p \left(\frac{c_{j,p}^{(h,AB)} - c_{j,p}^{(h,A)}}{c_{j,p}^{(h,AB)}} \right)^2},$$

où n_j est le nombre d'observations pour le logement j. Testé sur un ensemble de 676 logements, l'erreur quadratique moyenne relative δ_j est de 16,6% avec un écart-type de 13,9%.

Cette façon de valider le modèle bien qu'imparfaite puisque nous n'avons pas vraiment de vérité terrain, permet de s'assurer de la stabilité des distributions a posteriori inférées par l'algorithme sur des périodes différentes. L'erreur quadratique moyenne trouvée sur l'ensemble des 676 logements valide la stabilité du modèle.

Quant aux parts de chauffage inférées par l'algorithme, c'est le modèle physique du logement présenté section 2.2 qui permet d'assurer la validité des prédictions réalisées.

4.2 | Ajustement du modèle

L'ajustement du modèle dans le cadre de l'inférence variationnelle est une tâche assez complexe à cause d'une part de la stochasticité de la fonction objectif et du calcul de son gradient (voir 3.3) et d'une autre à cause des nombreux minima locaux dans lesquels l'optimisation peut se terminer si l'on ne prête pas assez attention à l'initialisation des paramètres.

4.2.1 | Difficultés rencontrées

La difficulté majeure rencontrée a été l'initialisation des paramètres. En les choisissant de façon maladroite l'optimisation se termine facilement dans in minimum local qui est bien loin de l'optimum global en terme d'explication des données par le modèle. Pour éviter une initialisation des paramètres trop loin de l'optimum, on utilise les pentes et biais des régressions linéaires sur les données pour $T < 17°C$ et $T \geq 17°C$, $17°C$ étant le maximum de la densité de probabilité des températures de cassure observées sur 676 logements.

4.2.2 | Critère d'arrêt

Afin d'ajuster les données de la meilleure façon possible en un minimum de temps, nous utilisons un critère d'arrêt qui se base sur le test de racine unitaire de l'ELBO et la

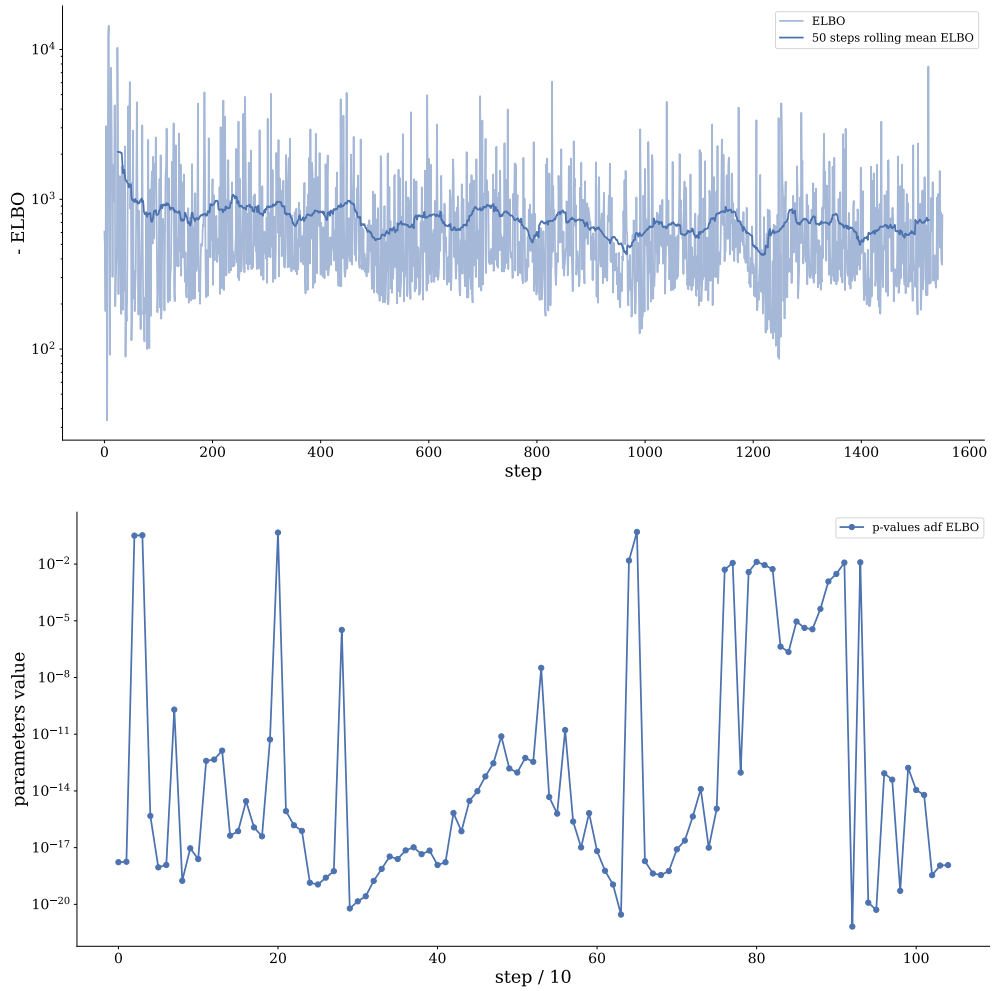

FIGURE 4.2 – Critère d'arrêt

variation relative des paramètres au cours des n derniers pas de calculs de l'optimisation. A cause du caractère stochastique du calcul de l'ELBO et de son gradient, sa trace au cours de l'optimisation a un comportement erratique (voir figure 4.2). Le test de racine unitaire vise à savoir si l'on peut considérer la série comme stationnaire. Si c'est le cas, on admet que la fonction objectif n'a plus de tendance baissière, signe que l'optimisation a convergé. Le test de Dickey–Fuller ayant une puissance statistique assez faible (rejet à tord de l'hypothèse de racine unitaire lorsque la racine est proche de 1) on ajoute à ce critère celui de variation des paramètres au cours des derniers pas. Ces deux critères permettent d'arrêter dans la majorité des cas l'optimisation lorsque l'optimum est atteint.

L'évolution des paramètres lors de l'optimisation de la figure 4.2 est tracée figure 4.3.

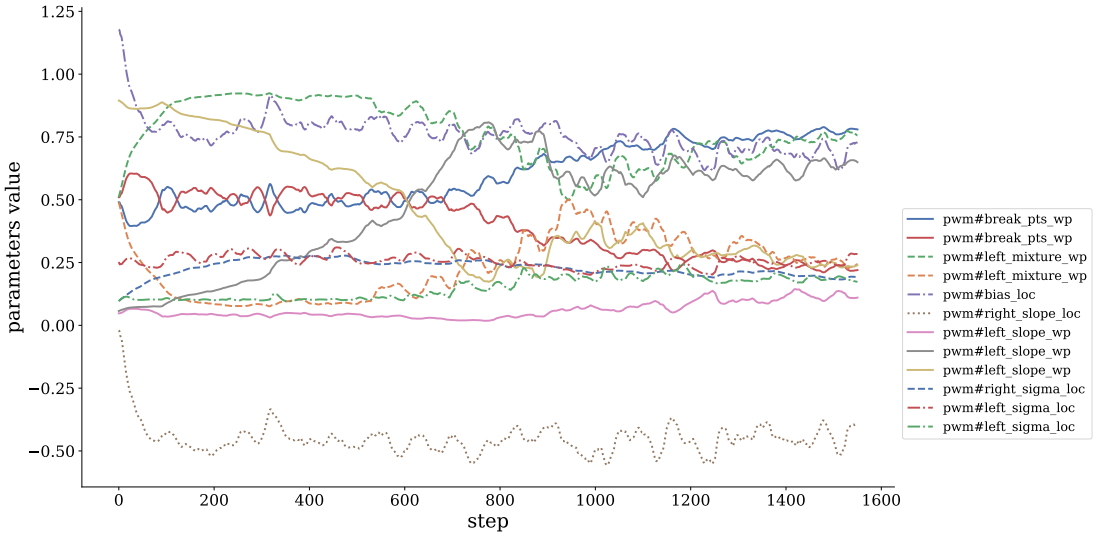

FIGURE 4.3 – Évolution des paramètres lors de l'optimisation

4.3 | Mise en production

Une grande attention a été portée sur la mise en production du nouvel algorithme pour qu'il soit robuste aux données manquantes et aux comportements étranges de consommation. On souhaite que l'algorithme puisse estimer le chauffage des 19 000 logements enregistrés chez Hello Watt en décembre 2020.

En effet, pour fonctionner correctement, le modèle à besoin d'assez de données de consommation pour des températures froides et chaudes sans lesquelles l'estimation de HTC n'est pas possible. De plus, il faut veiller à ce que les données de consommation aient une variance suffisante. On vérifie ainsi que les données de consommation ne soient pas nulles la majeure partie du temps, se qui correspondrait à une résidence secondaire par exemple.

4.3.1 | Pipeline d'inférence proposée

La pipeline d'inférence proposée (figure 4.4) permet de :

1. Tester les données d'entrée.
2. Ajuster le modèle le plus adéquat en fonction des données.
3. Enregistrer les paramètres du modèles pour les futurs entraînements du modèle et la prédiction du chauffage.

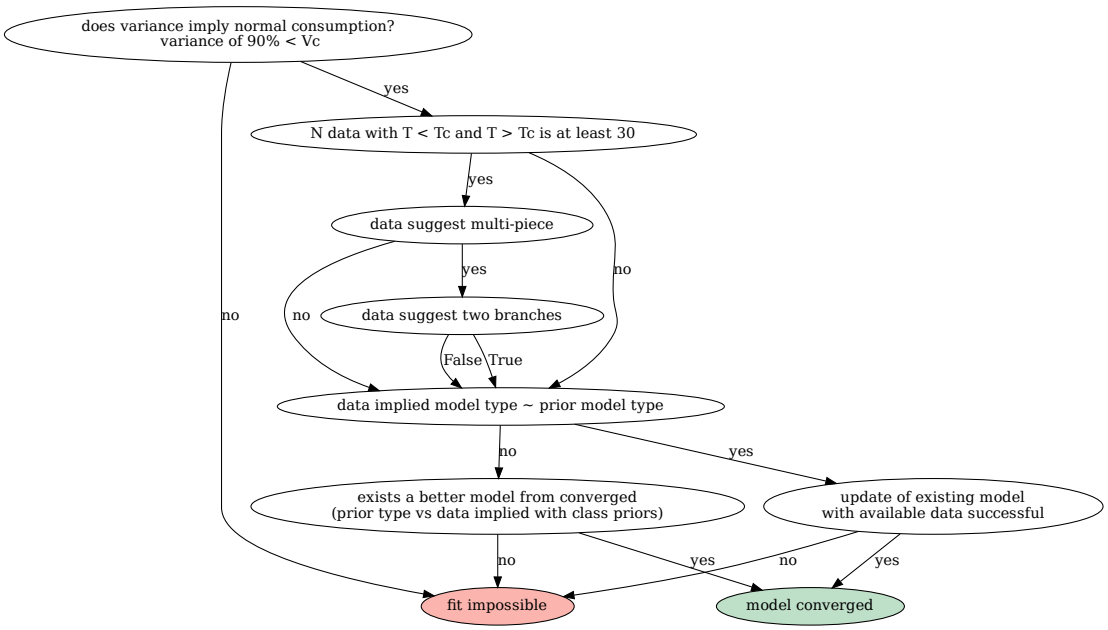

FIGURE 4.4 – Pipeline d'inférence

Le deuxième point de l'inférence proposé suggère le choix d'un modèle. Le modèle proposé, comme présenté dans la partie 3.4.2 ne se limite effectivement pas au modèle présenté dans ce rapport mais permet d'avoir n'importe quel nombre de points de cassure et d'états latents. On souhaite choisir le nombre de cassures et d'états latents pour avoir le modèle le plus simple qui explique le mieux les données. On peut utiliser le critère d'information bayésien (BIC) pour choisir ces paramètres. On rappelle que :

$$\text{BIC} = -2\ln(L) + k \cdot \ln(N)$$

$$\begin{aligned} L &= \text{vraisemblance du modèle,} \\ k &= \text{nombre de paramètres libres du modèle,} \\ N &= \text{nombre d'observations.} \end{aligned}$$

Cette façon de choisir les paramètres augmente les ressources en calcul nécessaires pour l'inférence puisqu'elle requiert d'ajuster beaucoup de configurations différentes pour chaque maison. La solution retenue est alors de choisir le modèle en fonction du nombre de données d'entrée et des critères statistiques sur la consommation pour choisir le nombre de cassures et d'états latents. Par exemple, pour choisir le nombre

d'états latents, on réalise une estimation de densité par noyau (KDE) sur les données de consommation des températures inférieures à $17°C$ avec une fenêtre fixe. On compte ensuite le nombre de modes de cette estimation de densité. Repérer deux modes distincts est un signe que la consommation sur cette plage de température est bimodale. On choisit alors d'ajuster le modèle avec deux états latents.

4.3.2 | Enregistrement des paramètres

Les paramètres d'inférence du modèle sont enregistrés à la fin de la pipeline pour :

1. Calculer les prédictions de chauffage lorsqu'un utilisateur se connecte sur le coach conso.
2. Utiliser ces paramètres pour les distributions a priori lorsque le modèle est entraîné à nouveau.

Le modèle est entraîné à intervalle de temps régulier pour vérifier que ses paramètres n'ont pas changé. Utiliser des paramètres erronés biaiserait les prédictions de chauffage. Nous citons ici deux cas pour lesquels ces paramètres viendraient à évoluer :

1. Les habitants du logement modifient leurs habitudes de chauffage. Par exemple, ils décident de chauffer lorsque la température extérieure moyenne de la journée descend en dessous de $15°C$ au lieu de $17°C$ précédemment.
2. Les résident ont réalisé des travaux d'isolation. Les pertes thermiques et le HTC de la maison sont alors modifiés.

L'enregistrement des paramètres demande ici de transformer les paramètres des distributions variationnelles en paramètres équivalents pour le modèle. Cette étape est triviale lorsque la distribution variationnelle choisie est la même que celle du modèle. Dans le cas contraire, on doit trouver les paramètres équivalents entre les deux distributions. On souhaiterait dans l'idéal avoir des distributions proches au sens de la divergence KL. Pour avoir une solution analytique, on peut utiliser l'espérance mathématique et la variance comme descripteurs des distributions et trouver les paramètres qui approchent ces deux quantités. Une solution pour l'approximation d'une loi Normale en une loi de Dirichlet est proposée en annexe B.

4.4 | Proposition d'amélioration

Notre approche de la modélisation graphique de la consommation électrique nous permet déjà de décrire des propriétés statistiques liées à la qualité d'isolation pour des

catégories de logements (telles que les logements regroupés par années de construction) et d'identifier des valeurs aberrantes dans ces catégories. On souhaite aussi par la suite détecter de façon non supervisé les logements qui utilisent le chauffage électrique. Ces logements sont pour le moment uniquement identifiés avec un formulaire remplit lors de l'inscription des utilisateurs sur le coach conso.

Nous envisageons également une approche plus générale de la modélisation bayésienne : plutôt que d'utiliser un seul modèle pour ajuster tous les cas possibles de données de consommation en faisant varier le nombre de points de cassure et d'états latents, nous souhaitons spécifier un ensemble de modèles pour en former des mixtures. On affecte alors des poids w_i pour chaque modèle et on souhaite trouver les poids de la mixture et les paramètres optimaux de chaque modèle qui explique le mieux les données, tout en pénalisant la complexité du modèle final.

Nous souhaitons aussi améliorer le modèle en ajoutant des données météorologiques pertinentes. On souhaite par exemple inclure dans le modèle la vitesse et la direction du vent, ainsi que les observations de la couverture nuageuse, en fonction de leurs importances.

5

Conclusion

Le stage réalisé au sein d'Hello Watt ne pouvait être qu'une expérience aussi exaltante qu'enrichissante. La qualité des équipes, l'esprit d'innovation et l'incitation à l'autonomie crée immanquablement une émulation dans l'ardeur au travail essentiel au bien être dans l'entreprise.

5.1 | Les objectifs accomplis

D'un point de vue de la prestation réalisée au cours de mon stage chez Hello Watt, j'ai développé un nouveau modèle non supervisé de désagrégation du chauffage, puis contribué à sa mise en production. Ces travaux ont par ailleurs fait l'objet d'une présentation lors du workshop international 2020 de la communauté NILM. [1] [2] Je suis coauteur d'un article qui va être publié dans les proceedings de cet évènement [3].

Le modèle développé augmente significativement la précision des prédictions de chauffage calculées précédemment avec un modèle supervisé entraîné à partir des données d'une maison de la base de donnée DRED. [4]

D'un point de vue technique, le modèle de désagrégation est un modèle bayésien implémenté avec la librairie de programmation probabiliste Pyro. Les paramètres sont ajustés pour chaque maison requérant une attention particulière lors de la mise en production sur le site d'Hello Watt. La pipeline proposée permet de répondre à ces exigences en vérifiant systématiquement les données d'entrée et en réalisant la phase d'apprentissage en amont du calcul des prédictions lorsqu'un utilisateur se connecte sur le site.

1. Lien vers la publication : http://nilmworkshop.org/2020/proceedings/nilm20-final23.pdf
2. lien vers la présentation : http://www.youtube.com/watch?v=0y75gJE_Eq4&t=290m47s
3. Lien arXiv : https://arxiv.org/abs/2011.05674
4. Lien : http://www.st.ewi.tudelft.nl/akshay/dred/

5.2 | Discussion

Les missions qui m'ont été confiées pendant le stage sont révélatrices des responsabilités qui affèrent au data scientist dans une organisation. L'annexe C présente un exemple d'analyse statistique qui m'a été confiée permettant de présenter sur le site internet d'Hello Watt la répartition des étiquettes DPE de chaque département en France en retirant le bias d'estimation lié à l'échantillonnage de la base de données de l'agence de la transition écologique (ADEME). Cet exemple illustre la variété des tâches qu'un data scientist doit être capable de traîter dans une organisation. Dans le monde des entreprises actuel, rapide et hautement compétitif, la gouvernance des données est indispensable. Le métier du data scientist est dans ce contexte non réduit à la création de modèle mais s'étend aussi à la gestion et l'exploitation des sources de données externes et internes des entreprises.

A

Inférence bayésienne de la moyenne d'une Gaussienne

L'inférence bayésienne de la moyenne d'une Gaussienne ayant un écart type connu à l'advantage d'avoir une solution analytique lorque la distribution à priori est aussi une Gaussienne. Les distributions à postériori et à priori sont alors conjuguées.

Sous les hypothèses choisies, la distribution à priori s'écrit :

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}}$$

La vraisemblance est lié à notre modèle : connaissant la moyenne inconnue, la distribution des données est une Gaussienne.

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

La distribution à postériori est alors :

$$f(\mu \mid x) = N(\mu_1, \sigma_1)$$

Avec :

$$\sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \right)^{-1}$$

$$\mu_1 = \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n} \right)$$

B

Approximation d'une loi normale en une loi de Dirichlet

On souhaite trouver les paramètres d'une loi de Dirichlet qui approchent le mieux une loi Normale au sens de la moyenne et de la variance, faute de pouvoir analytiquement trouver les paramètres optimaux au sens de la divergence KL. Les résultats qui suivent supposent que la loi de Dirichlet n'est pas creuse ($\alpha_i \ggg 1$). On note $\alpha_0 = \sum \alpha_i$.

Les propriétés utilisées de la loi de dirichlet sont :

$$\begin{cases} E[X_i] &= \frac{\alpha_i}{\alpha_0} \\ V[X_i] &= \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)} \end{cases}$$

En dimension 1 La loi de Dirichlet est alors mieux connue sous le nom de loi Beta. L'égalité des moyennes et des variances s'écrivent :

$$\begin{cases} \mu &= \frac{\alpha}{\alpha+\beta} \\ \sigma^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{cases}$$

En utilisant l'approximation :

$$\alpha + \beta + 1 \approx \alpha + \beta$$

$$\begin{cases} \alpha &= \frac{(1-\mu)\mu^2}{\sigma^2} \\ \beta &= \alpha\frac{1-\mu}{\mu} \end{cases}$$

En dimension 2 L'égalité des moyennes et des variances s'écrivent :

$$\begin{cases} \mu_1 &= \frac{\alpha_1}{\alpha_0} \ (1) \\ \mu_2 &= \frac{\alpha_2}{\alpha_0} \ (2) \\ \sigma_1^2 &= \frac{\alpha_1(\alpha_2+\alpha_3)}{(\alpha_0)^2(\alpha_0+1)} \ (3) \\ \sigma_2^2 &= \frac{\alpha_2(\alpha_1+\alpha_3)}{(\alpha_0)^2(\alpha_0+1)} \ (4) \end{cases}$$

$$\alpha_0 + 1 \approx \alpha_0 \ (5)$$

$$\begin{cases} (1),(3),(5) \Rightarrow \sigma_1^2 = \frac{\alpha_2+\alpha_3}{\alpha_1^2}\mu_1^3 = \frac{(1-\mu_1)\mu_1^2}{\alpha_1} \\ (2),(4),(5) \Rightarrow \sigma_2^2 = \frac{\alpha_1+\alpha_3}{\alpha_2^2}\mu_2^3 = \frac{(1-\mu_2)\mu_2^2}{\alpha_2} \end{cases}$$

$$\begin{cases} \alpha_1 = \frac{(1-\mu_1)\mu_1^2}{\sigma_1^2} \\ \alpha_2 = \frac{(1-\mu_2)\mu_2^2}{\sigma_2^2} \text{ ou } \alpha_2 = \frac{\mu_2}{\mu_1}\alpha_1 \\ \alpha_3 = \alpha_1\frac{1-\mu_1}{\mu_1} - \alpha_2 \end{cases}$$

Puisque la loi de dirichlet n'a que 3 paramètres contres 4 pour la loi Normale, le problème est sur-contraint et une des égalités doit être levée. On choisira $\alpha_2 = \frac{\mu_2}{\mu_1}\alpha_1$ si on souhaite préserver l'égalité des moyennes.

C

Post stratification de la répartition des étiquettes DPE

Le jeu de données de l'ADEME contient les données des maisons évaluées pour les étiquettes DPE. L'analyse statistique de ces données à deux limitations :

1. L'évaluation des DPE est seulement réalisée au moment de la vente ou de la mise en location d'un bien immobilier. Ainsi, les maisons récentes sont sur-représentées dans le jeu de données de l'ADEME et le nombre d'étiquettes A et B est sur-estimé à cause de la corrélation entre l'année de construction d'un bâtiment et sa consommation en énergie.
2. La méthode d'évaluation des bâtiments n'est pas aussi bonne pour toutes les catégories. Pour les maisons construites avant 1948, l'évaluation peut être conduite sur la base des factures d'électricité des ménages. Cette méthode amène un bias provenant des différences de comportement sur l'utilisation du chauffage des ménages, indépendamment de la qualité d'isolation de leurs logements. Encore une fois, ce bias augmente artificiellement le nombre d'étiquettes A et B et diminue le nombre d'étiquettes F et G.

Pour limiter ces bias, on peut associer les données de l'ADEME à celles du recensement de l'INSEE de cette façon :

1. Calculer $P(\text{DPE} \mid \text{âge bâtiment, code INSEE})$ avec les données de l'ADEME,
2. Calculer $P(\text{âge, code INSEE})$ avec les données de l'INSEE,
3. Calculer $P(\text{DPE, code INSEE}) = \sum P(\text{DPE} \mid \text{âge, code INSEE}) P(\text{âge, code INSEE})$.

Cette méthode appelée post stratification permet de réduire le bias d'un estimateur statistique provenant d'une population mal échantillonnée.

Bibliographie

[1] Benjamin J. Birt, Guy R. Newsham, Ian Beausoleil-Morrison, Marianne M. Armstrong, Neil Saldanha, and Ian H. Rowlands. Disaggregating categories of electrical energy end-use from whole-house hourly data. *Energy and Buildings*, 50 :93–102, June 2012.

[2] Amir Kavousian, Ram Rajagopal, and Martin Fischer. Determinants of residential electricity consumption : Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, 55 :184–194, 2013.

[3] Stephan Spiegel and Sahin Albayrak. Energy disaggregation meets heating control. In *Proceedings of the 29th ACM Symposium on Applied Computing*, SAC '14, page 559–566, New York, NY, USA, March 2014. ACM.

[4] Yiling Jia, Nipun Batra, Hongning Wang, and Kamin Whitehouse. A tree-structured neural network model for household energy breakdown. In *The World Wide Web Conference*, WWW '19, pages 2872–2878, New York, NY, USA, 2019. ACM.

[5] Haroon Rashid, Nipun Batra, and Pushpendra Singh. Rimor : Towards identifying anomalous appliances in buildings. In *Proceedings of the 5th Conference on Systems for Built Environments*, BuildSys '18, pages 33–42, New York, NY, USA, 2018. ACM.

[6] Haroon Rashid, Pushpendra Singh, Vladimir Stankovic, and Lina Stankovic. Can non-intrusive load monitoring be used for identifying an appliance's anomalous behaviour ? *Applied Energy*, 238 :796–805, March 2019.

[7] Jonathan D. Chambers. *Developing a rapid, scalable method of thermal characterisation for UK dwellings using smart meter data*. PhD thesis, University College London, December 2017.

[8] Panagiota Gianniou, Christoph Reinhart, David Hsu, Alfred Heller, and Carsten Rode. Estimation of temperature setpoints and heat transfer coefficients among residential buildings in denmark based on smart meter data. *Building and Environment*, 139 :125–133, July 2018.

[9] Chirag Deb, Mario Frei, Johannes Hofer, and Arno Schlueter. Automated load disaggregation for residences with electrical resistance heating. *Energy and Buildings*, 182 :61–74, October 2018.

[10] Srinivasan Iyengar, Stephen Lee, David Irwin, Prashant Shenoy, and Benjamin Weil. Watthome : A data-driven approach for energy efficiency analytics at city-scale. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 396–405, New York, NY, USA, 2018. ACM.

[11] Jonathan D. Chambers and Tadj Oreszczyn. Deconstruct : A scalable method of as-built heat power loss coefficient inference for uk dwellings using smart meter data. *Energy and Buildings*, 183 :443–453, January 2019.

[12] Valentin Roussel. *Introduction à l'inférence bayésienne*. HAL, 2019.

[13] Scott M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, 1007.